DOCUMENT RESUME

ED 152 690 95 SP 012 208

AUTHOR Penner, William J.; Gilmore, Joseph T.
TITLE CBTE in Special Education: Competency Assessment.
INSTITUTION New York State Education Dept., Albany. Office for the Education of Children with Handicapped Conditions.; State Univ. of New York, Plattsburgh. Coll. at Plattsburgh.
SPONS AGENCY Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C. Div. of Personnel Preparation.
PUB DATE Aug 76
NOTE 177p.; Proceedings of a conference held at Plattsburgh, New York, August 10-13, 1976

EDRS PRICE MF-$0.83 HC-$10.03 Plus Postage.
DESCRIPTORS Conference Reports; Data Analysis; Data Collection; Higher Education; *Instrumentation; *Performance Based Teacher Education; *Preservice Education; Reliability; *Special Education; Systems Development; *Teacher Evaluation; Validity

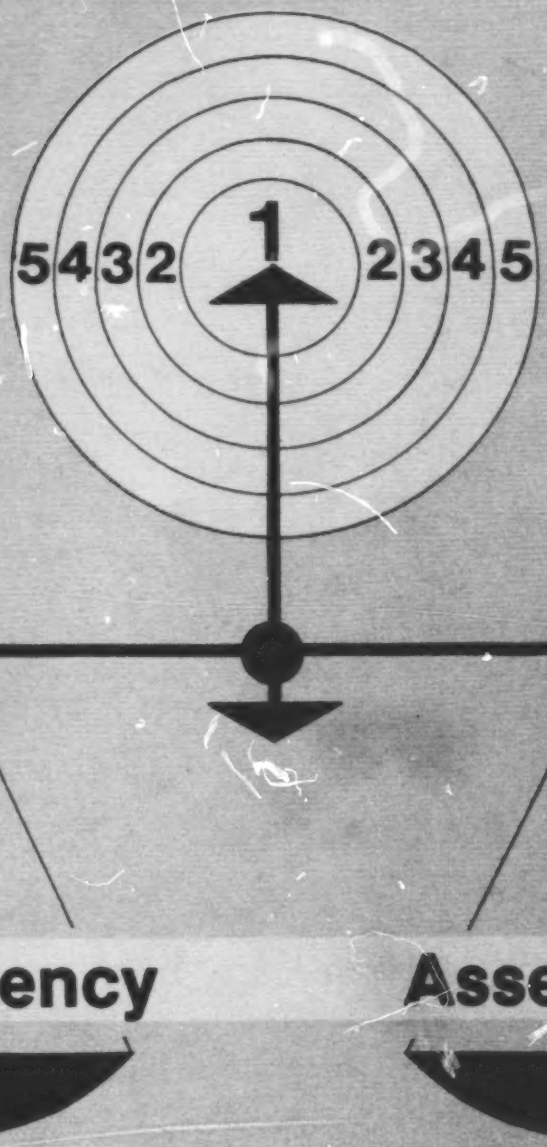
ABSTRACT

This document presents proceedings of a 1976 conference on the assessment of teaching competence in special education teachers. The concerns identified for discussion were: (1) instruments for assessing interpersonal competencies in preservice special education students; (2) instruments for assessing teaching competencies in noninterpersonal areas; (3) ways and means of insuring reliability and validity in competency assessment; and (4) designing systems for managing competency assessment and resulting data. The document presents, for each of these four objectives, a resource paper presented by a conference speaker, transcript of the conference session devoted to the paper, and a reply to the content of the paper. (HJB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED152690

CBTE IN SPECIAL EDUCATION:



U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PROCEEDINGS OF A CONFERENCE HELD AT PLATTSBURGH, NEW YORK, AUGUST 10-13, 1976

William J. Penner
Joseph T. Gilmore

SP012209

CBTE IN SPECIAL EDUCATION:
COMPETENCY ASSESSMENT

Edited by

William J. Penner

Joseph T. Gilmore

PROCEEDINGS OF A CONFERENCE HELD IN
PLATTSBURGH, NEW YORK, AUGUST 1976

Sponsored by

The University of the State of New York
The State Education Department
Office for the Education of
Children with Handicapping Conditions

In cooperation with

The Special Education CBTE Project
State University of New York
College of Arts and Sciences
Plattsburgh, New York 12901

This document has been made possible by funding awarded to the Research Fund of the Board of Regents of the University of the State of New York, the State Education Department, Office for the Education of Children with Handicapping Conditions, by the United States Office of Education, Bureau of Education for the Handicapped, Division of Personnel Preparation, Grant Number G007602968, Project Number 451AH60852.

Copies of these proceedings may be obtained from:

Special Education Program
State University of New York
College of Arts and Science
Plattsburgh, NY 12901

Office for the Education of
Children with Handicapping Conditions
State Education Department
55 Elk Street - Room 115
Albany, NY 12234

CONTENTS

Preface v

Acknowledgements viii

INTRODUCTION William J. Penner 1

OBJECTIVE 1: To develop or improve existing instruments for measuring student competencies other than interpersonal skills --

Resource Paper
Developing and Improving Instruments for Measuring the Competence of Preservice Teacher Education Students
Gilbert Shearron 5

Conference Session
Gilbert Shearron 21

A Reply to Dr. Shearron's Paper
James D. Strauch 25

OBJECTIVE 2: To develop instruments to more adequately assess interpersonal skills of students --

Resource Paper
Techniques and Criteria for Designing and Selecting Instruments for Assessing Teachers
Paul Pottinger 28

Conference Session
Paul Pottinger 73

In Response to the Paper by Paul Pottinger
Joyce Spence 79

OBJECTIVE 3: To specify ways and means of insuring reliability and validity in competency assessment --

Resource Paper
Specifying Ways and Means of Insuring Reliability and Validity in Competency Assessment
Beatrice A. Ward and
William J. Tikunoff 84

Conference Session
Beatrice A. Ward and
William J. Tikunoff 102

Some Random Remarks About Ways and Means of Insuring Reliability and Validity in Competency Assessment
Patricia M. Kay 109

OBJECTIVE 4: To design a system or components of a system for managing the assessment program and resultant data --

Resource Paper
Designing a System for Managing Assessment in Competency-Based Teacher Education Programs: A Worktext
J. Hugh Baird 115

Conference Session
J. Hugh Baird 141

CONFERENCE WRAP-UP

Competency Assessment: A Perspective
Robert Brinkerhoff 145

APPENDICES

A. Needs Assessment Survey 157

B. Conference Program 160

C. Conference Consultants 162

D. Conference Participants 163

E. Conference Planning and Management Committee 165

F. Conference Evaluation 166

PREFACE

One of the earliest recorded observations of systematic philosophy is that of Heraclitus (540 B.C.) who said, "All things flow". There are modern day thinkers who would agree with the Greek philosopher that change is the only reality, that the only constant is change and that this is true on both the individual and societal levels. It is in the context of change that competence based teacher education, like other trends and movements in society, must be viewed. It is not to our purpose here to attempt to delineate the historical antecedents and social forces that may have contributed to the development of this strong movement in recent educational history or to predict its eventual outcome. It may be helpful background, however, to trace briefly the genesis of this document.

In June, 1972, the Research Fund of the Board of Regents of the University of the State of New York, through the Division for Handicapped Children of the State Education Department was awarded a grant to explore the competency based hypothesis and its implications for teacher training, particularly with regard to providing alternative training strategies and models for the certification of teachers (The CBTE Project in Special Education). Funding support was provided by the Division of Personnel Preparation, Bureau for Education of the Handicapped, U.S. Office of Education. The original objectives of the project were to support CBTE research and development underpinnings in that context. An action of the Board of Regents in September, 1972 gave great visibility and immediate relevancy to the work of the CBTE project. In September, 1972, the Board of Regents endorsed a competence based, field centered approach to the professional preparation of educational personnel, effectively mandating that college and university teacher preparation facilities actively pursue a competence based approach in the professional training and educating of prospective public school personnel.

New York State defines a competence based program of preparation as one which provides acceptable evidence of program formulation through the collaborative efforts of a consortium of representatives of colleges and universities, school district administration, and professional staff of school districts. So defined, a competence based system permits the widest variety of program design since it does not prescribe any set of course or learning activities. A schedule established by the State Education Department's Division of Teacher Education and Certification called for all college and universities preparing elementary and special education personnel to submit competence based training program proposal plans by February 1, 1975 to be reviewed for program registration.

The Regents mandates facilitated the CBTE project policy board decision to widen the horizons of the CBTE project. The acquisition, development and evaluation of instructional materials and assessment techniques continued to be primary objectives of this research and development project. It had never been the intention of the CBTE project in special education to "sell" CBTE as the most viable alternative for teacher education programs. However promising the prospects of competence based teacher education are viewed or whatever serious deficiencies CBTE may have, the primary objective continued to be to explore the competence based hypothesis, but the main focus shifted to

allow for providing technical assistance to the consortia of forty-four institutions of higher education around New York State wrestling with the collaborative process of designing preparatory programs in special education.

With emphasis on assistance to the consortia in the development of their programs, the project sponsored, among other activities, three statewide conferences to which were invited the forty-four consortia. The proceedings of the first conference held in Glenmont, New York on May 14 through 16, 1974, was published under the title, Design for Competence Based Education in Special Education. The proceedings of the second conference held in Glenmont, New York, October 29 through 31, 1974, have been printed under the title, Assessment in CBTE: The Search for Evidence. The present document furnishes the proceedings of the third statewide conference and focuses on CBTE in Special Education: Competency Assessment. The CBTE Project is grateful to Dr. William J. Penner, Coordinator, Special Education Program, State University College at Plattsburgh, for his efforts in arranging the third statewide conference as a cooperative venture with the State Education Department, Office for Education of Children with Handicapping Conditions.

In a final report for the CBTE Project in Special Education may be found reflections at greater length on the products and processes of the various project activities. The point may be made briefly here, however, that there was a good bit of evident change in the substance and style between the first and the third conferences. Each conference served a timely need, but the contextual referents were quite different. The Plattsburgh conference, held almost a year and a half after the preparatory programs had submitted their program proposals for approval and registration by the State Education Department, Division of Teacher Education and Certification, drew the smallest number of participants and the widest degree of acceptance by the participants. Some other significant variables certainly may have been the holding of the conference in late summer and/or the particular talents of the consultants, and/or the greater availability of the consultants for individual attention to the fewer participants. Whatever the variety of reasons, there seemed to be consensus on the part of the participants that they had achieved significant progress towards a gestalt in the area of competency assessment. This perception of the participants at the Plattsburgh conference was unique in its near universality and intensity.

No suggestion is made that anything approaching a gestalt has actually been achieved, either as a result of conferences or whatever. With regard to the whole movement in competence based teacher education, we still find ourselves each approaching the total task of teacher training and teacher competency much as the proverbial series of blind men and the elephant. It is not a sense of smugness but more likely a humble view of reality to suggest that the state of the art in competence based teacher education still allows for a variety of approaches to the central issues the movement has raised. Teacher education, like any discipline, has its "secret agents": those who have the secret answer to complex issues. The state of the art suggests otherwise. The building of a body of literature and sound research conclusions goes on. Pablo Neruda may be to the point in this context, "We have to discard the past and as one builds floor by floor, window by window, and the building rises, so do we go on throwing down first broken tiles, then

pompous doors..." The tearing down and building up process goes on simultaneously, and generally at a far slower pace than enthusiasts are willing to admit.

The editors of each of the proceedings hope they have been builders. The editors of this document are pleased to present this proceedings document to readers involved in the daily flux enveloping teacher education in general and special education in particular. There is a sector approach in the area of psychotherapy which believes that understanding and healing a part of a problem can contribute to a deeper understanding and pervasive healing of the whole. It is hoped that careful study of the engaging and creative approaches of the consultants whose contributions are contained in this document may contribute to each reader's growth towards a closer approximation of your own gestalt in the area of teacher education through a better understanding of competency assessment.

Joseph T. Gilmore
Albany, New York
September, 1977

ACKNOWLEDGEMENTS

Project personnel gratefully acknowledge the assistance and encouragement of numerous persons in planning and conducting the conference:

The CBTE project in special education has had funding support and technical assistance from the Division of Personnel Preparation, Bureau of Education for the Handicapped, U.S. Office of Education, through Dr. Thomas Behrens, Project Officer.

Particular thanks is extended to the New York State Special Education CBTE Policy Board for perspectives on competency assessment as viewed by teacher organizations, school administrators, teacher educators and the State Education Department.

Gratitude also to college and university consortia who articulated their competency assessment needs.

The conference consultants were most cordial in their responsiveness to participant concerns.

Patricia Kay, City University of New York, and Theodore Andrews of the Concern for Educational Development, Reston, Virginia, contributed significantly to the conference process and the preparation of this report.

Dale Nitzschke, Associate Dean of Professional and General Studies, and Raymond LaRocque, Coordinator of Educational Conferences, SUC at Plattsburgh, were very helpful in arranging for fine facilities, meals, lodging and registration.

Carol Bleaux, Sandra Kimble and Christine Rotella provided gracious secretarial support.

INTRODUCTION

William J. Penner
State University College
Plattsburgh, New York

Role of the CBTE Project in Special Education

In the preface to these proceedings, Joseph T. Gilmore has noted the standing commitment of New York State to the improvement of teacher education and to improved accountability for teacher education outcomes via competency-based teacher education (CBTE) models. He has further pointed out the role of the CBTE Project in Special Education in exploring with Special Education CBTE consortia the resources, processes and needs inherent in such undertakings, as well as project efforts to provide technical assistance to facilitate consortium development of CBTE programs in Special Education.

Conference Need

As programs were developed and implemented, questions regarding the assessment of teaching competence became more numerous, more specific and more probing. These competency assessment questions deal with such basic concerns as validity, clarity and specificity of the competencies to be assessed; and with the validity and reliability of the assessment techniques and instruments employed. Related to such broad concerns, but of a more logistical nature, are questions concerning which or how many competencies to assess, how often, by whom; and how to manage, interpret and utilize resultant data.

Conference Objectives

Confronted with this vast array of questions, problems and needs; and faced with the need to limit the scope of the conference, it was apparent that a three-day conference could attempt to respond to only selected aspects of competency assessment. Moreover, it was felt that the process of selecting those aspects to be addressed by the conference ought to draw planning input from the consortia and other parties concerned with CBTE. Hence, project personnel, in consultation with the State CBTE Policy Board, developed a Needs Assessment Survey (Appendix A) in order to identify the major competency assessment concerns of Special Education CBTE consortia. A tabulation of survey responses indicated wide-spread and high priority concern for improving the assessment of competencies to be demonstrated in practicum settings. The particular aspects of competency assessment most frequently noted as consortium concerns became the basis of the following conference objectives:

1. Instrumentation for assessing interpersonal competencies of Special Education students

2. Instrumentation for assessing teaching competencies in non-interpersonal areas

3. Specifying ways and means of insuring reliability and validity in competency assessment

4. Designing systems for managing competency assessment and resultant data

Conference Format

Also indicated in the needs assessment survey was the desire of responding consortia to spend a portion of conference time in hands-on workshop activities in which the technical assistance of conference consultants might be applied to specific competency assessment concerns of the respective consortia. Thus, the conference format addressed each objective in turn with:

1. Participant reading of a resource paper prepared by a consultant;

2. A presentation by that consultant, followed by questions from the floor;

3. A workshop session in which participants could apply the material presented to their respective competency assessment needs with the assistance of consultants and facilitators;

4. The presentation by a second conference consultant of a response to material presented and/or of additional perspectives of relevance to the objective in question.

The conference program appears in Appendix B.

Conference Consultants

The planning committee was indeed fortunate in retaining consultants (listed in Appendix C) with a high degree of experience and expertise pertaining to conference objectives. In addition to providing technical resources and assistance, consultants met daily with the planning committee to review the day's events. Their flexibility in responding to the immediate questions and concerns expressed by participants is viewed as a key factor in maintaining conference relevance.

Conference Content

Consultants join in expressing their shared concern about the continuing need to more clearly and operationally define teaching competence even while striving to develop means of assessing as accurately as possible the presence of competencies as stated in existing CBTE programs. There are reservations springing from the presumptuousness of assessing

what is understood only in part; but conference spirit also conveys a strongly proactive orientation toward the evolution of more adequate concepts of competence and its measurement. As the proceedings unfold, the tension between pragmatic action and challenging idealism is both stimulated and relieved by the wide assortment of consultant contributions.

In the initial workshop session, Dr. Shearron deals with means of developing or improving instrumentation for competency assessment, and highlights a number of concerns that recur throughout the conference. After recommending attention to the ways in which teacher education may be influenced by such contexts as on-campus study, simulation and work or practicum settings; he points out that different teaching competencies may be appropriate for unstructured classroom situations than for structured learning activities. A further dimension is the use of structured interview protocols to obtain context-relevant information that can be used to interpret the appropriateness of teaching behaviors that are observed and rated. The question of selecting those behaviors to be assessed is linked to questions involving the frequency of assessment, the need for multiple observers and whether the assessment is formative or summative in nature.

In responding to Dr. Shearron's presentation, Dr. Strauch emphasizes the need for a taxonomy of teaching competencies to facilitate the development of manageable evaluation plans and to foster understandable communication about CBTE programs. He notes a need to evaluate competence in work contexts; but indicates that logistical limitations underscore the importance of developing an approach to the sampling of trainee performance.

In dealing with assessment of interpersonal competence, Dr. Pottinger cautions that much valid and reliable measuring can be done without yielding meaningful data if there is an overuse of reductionistic, low-inference measures. In order to assess teaching competence, he builds a strong case for assessing trainees while they are operating on the environment, and for assessing such causally related variables as warmth, flexibility and the ability to work under stress. Dr. Pottinger's theses are supported with reference to competency assessment studies that illustrate methods for applying rigorous measurement techniques to relevant causal and situational variables.

Sharing the session on assessment of interpersonal competence, Dr. Spence mentions the need for teacher educators to move beyond the consideration of teaching behaviors to their causes, particularly the importance of self-awareness of special education trainees and the relationship of self-awareness to job performance. She also notes with illustrations the importance of measuring in trainees the same affective outcomes that are desired for exceptional pupils, and suggests that appropriate criteria of competence might be based on individual trainee growth rather than on some fixed and universally expected level of performance.

Issues of reliability and validity of competency assessment are presented by Drs. Ward and Tikunoff. They note that if reliability and validity in the assessment of teaching competence are to be realized, different teaching acts must be viewed in terms of their functional equivalency in a variety of intra-teacher and inter-teacher contexts. They propose a matrix of context variables, including pupil characteristics, teacher characteristics and curricular variables. In addition, they illustrate ways in which multiple observers provide not only for the measurement of reliability, but for consideration of alternative educational values as a means of determining the validity of the teaching competencies being assessed. This multivariate approach to competency assessment is both challenging and gratifying in concept to those wishing to consider appropriate and effective competencies as flexible responses to infinitely complex and varied teaching situations.

Dr. Kay expands upon the use of a multivariate approach by pointing out that researchers, school administrators and teacher educators often have needs that call for varying hardness of data, and that teacher educators might well respect the accumulated, traditional wisdom of the profession while improving assessment reliability and validity by using limited, but generally available, resources. She offers reassuring guidelines for bridging the gap between teacher education needs (how we can assess competence) and continuing research efforts (what competence is).

A major concern of the conference addressed by Dr. Baird is management of a competency assessment system. He takes us step by step through the development of a competency assessment system with adherence to two important admonitions: (1) parsimony of assessment -- or, don't collect unimportant data; and (2) decision making should involve those affected. A further point of significant import is that the task can be accomplished without a large increase in program resources.

Dr. Brinkerhoff provides an overview of the entire competency assessment process. He underscores major emphases of the conference with practical suggestions regarding the use of existing assessment instruments and data on the grounds that competency assessment can improve once it is started.

In summary, these proceedings provide the expertise of multiple consultants regarding:

-- accounting for the effects of context variables,
-- identifying measures for variables that are internal, but causally or predictably related to competent teaching performance,
-- the linking of reliability and validity to situational variables and educational values,
-- the feasibility of developing competency assessment systems with the resources available.

It is hoped that the reader will be enriched by the concepts and techniques presented in these proceedings, and will share with conference participants, consultants and sponsors the conviction that competency assessment can be both manageable and relevant.

DEVELOPING AND IMPROVING INSTRUMENTS FOR MEASURING THE COMPETENCE OF PRESERVICE TEACHER EDUCATION STUDENTS

Gilbert Shearron
University of Georgia

The measurement of competence is perhaps the most difficult task faced by developers and implementers of Competency Based Teacher Education (CBTE) programs. This paper attempts to focus on only one phase of the assessment problem, the development and/or improvement of instruments that can be used in the measurement of the competence of preservice teacher education students. The paper does not deal with the affective area of teacher competence. The paper does not deal with the problems of researching teacher effectiveness nor does it deal with the philosophical issues of the desirability of a data base before proceeding with the measurement of competence.

The term measurement can be defined as a procedure where numbers are uniquely assigned to entities: persons, other organisms, objects, statements, etc., (Cureton, 1969). The entity we are concerned about is the measurement of the performance of people and the assignment of some sort of value to that performance. Assessment and evaluation are broader terms. Measurement is often considered as a part of assessment and evaluation.

The term competence can be defined as being qualified, capable, or adequate. The term competency is defined by Johnson, Shearron, Hensel, (1974) as being a rational performance which satisfactorily meets the objectives for a desired condition. There are many parts of a competency. Performance is an observable behavior, but it includes the manipulation of ideas and the making of judgements and decisions. This paper deals only with the measurement of the observable performance part of a competency.

Questions to be Considered

There are a number of questions that need to be considered before one makes decisions concerning the development and improvement of instruments for measuring the competence of preservice teacher education students. These questions are interrelated. Consideration of one without the consideration of the others will probably leave us with measurement devices that are not useful in the real world of a teacher education program.

1. At what level of specificity are the competencies that are to be measured stated? A major problem in measuring competence is the lack of operational definitions for the competencies to be measured. Instruments designed to measure competence should relate directly to the statements of competence. The more specific the competency statements,

the easier the task of measurement. Failure to operationally define competencies leaves us with highly subjective measures that are no better than what has traditionally been used. Consider the following competency statement found on many lists of competencies. The teacher gives clear directions to pupils. What does this mean in terms of measurement? The situations in which this competency might be measured could differ significantly. For example, the preservice teacher could be giving directions for behavior in the classroom or the directions could be the assignment of a lesson in mathematics. An operational definition is needed that defines specifically what kind and type directions are to be measured.

2. In what context is the demonstration of the competency to be measured? There are a number of contexts in which competencies can be measured. Turner (1974) identifies three which are found in many teacher education programs.

a. Symbolic Context. This refers to the college or university context. Typically, only verbal performance or what the preservice teacher knows as opposed to what he/she actually does.

b. Simulated Context. This is the laboratory context. Films, video tapes and other kinds of aids are employed. Teaching is restricted to such things as microteaching. The context is quasi-concrete, but verbal performance predominates.

c. Work Context. This is the "real" classroom setting where students have opportunities to integrate knowledge and action. Performance is what is measured in this context. Shearron and Johnson (1973) divide the work context into two parts. Part one is referred to as working with pupils in structured situations. This means that classroom situations are specifically designed for the student to practice and demonstrate specific teaching skills. For example, a social studies lesson is structured to be a question-answer situation to enable the student to demonstrate this particular skill. The second part of the work context is referred to as an unstructured situation. In an unstructured situation, preservice teachers are evaluated when they demonstrate competencies as the need for those competencies occur. An illustration of this is that the preservice teacher would use skills in classroom management as they are needed.

The question of context is extremely important. It is much easier to develop or adapt instruments to measure competence in a structured work context than it is to develop instruments that are all inclusive to be used in an unstructured work context. Attempts to do the latter have at least in the past often ended up as checklists that cover everything in a highly subjective manner.

Measurement in a symbolic context is the easiest and the most familiar to most teacher educators. Paper and pencil tests and other ways of measuring knowledge do not necessarily require the development of instruments. The instruments developed for measurement in a work context can often be adapted to the simulated context.

3. How often and at what points are preservice teachers to be measured? A major problem in measuring performance by observation in the unstructured work context is controlling the sampling of individual

behavior. McNeil and Popham (1973) point out that a major deterent to the use of rating scales is the failure to control for sampling teacher behavior. A strategy for sampling behavior must be developed along with the development of instrument designed to measure performance by observation.

4. Who is to be involved in measuring the performance of preservice teachers? In the preservice program there are several categories of individuals who could be involved in the measurement process. The question of context is, of course, closely related to this question. Assuming that measurement of competence is to take place in the work context, we would have college and university personnel, classroom teachers, and peers, available to measure the competence of the preservice student. In the structured work context any of these three categories of persons could be employed. However, in the unstructured work context it would be extremely costly to involve anyone other than the classroom teacher who would be there anyway.

5. Will the measurement of competence be designed to be formative, summative, or both? This is an extremely important question to be resolved. As we shall see later in this paper, formative evaluation, which in the author's view is part of a training program, lends itself to using systematic observation instruments that focus on bits and pieces of a teaching competency. While summative evaluation should consider the total competency, there are some types of measuring devices which are more appropriate for one than the other.

6. How much time will be required for measurement? A very real problem is the amount of time required to measure competence. In adapting or developing any instruments for measurement consideration must be given to the amount of time involved in utilizing the instrument. In general, the instruments will need to allow for the collection of data at a reasonable rate.

7. What is to be measured? This final question is one that cannot be answered by saying, "we are measuring competence." This is not an operational statement that can be translated into some type of action. The sub-questions that fall out of this more inclusive one are: Do you measure every student on every competence? Do you sample behavior in the unstructured work context? Do you limit measurement to the structured work context? These questions need to be dealt with prior to making decisions about the instruments needed to measure competence.

What Can We Use?

There are currently hundreds of instruments available to measure teacher behavior. Kay (1974) reports that these instruments fall roughly into two types, neither of which are wholly appropriate for measuring the performance of preservice teachers. The first type contains instruments used by school systems to rate personnel and instruments developed by colleges and universities to measure performance in student teaching. These instruments are very subjective because of the lack of operational definitions for the categories being measured by the instruments. The second type of available instruments contains those that have been constructed primarily for use in research. These instruments usually

focus on some highly defined and very specific teacher behavior. These instruments can provide evidence that preservice teachers possess specified skills in isolation, but they do not measure the complexities of classroom situations.

The criticism made by many that the instruments to measure performance are not available is to a great extent true. However, the American Association of Colleges for Teacher Education's Committee on Performance Based Teacher Education (1975) points out that the lack of instruments is clearly not unique to CBTE programs. The committee recommends that it is time to spend less time on talking about the lack of capability to measure teacher performance and to put our energies on increasing that capability.

The next section of the paper will focus on the types of instruments that can be used to measure competence. Attention will be given to four types of instruments.

Systematic Observation

Furst and Hill (1971) define systematic observation as a set of procedures which uses systems of categories to code and to quantify classroom behaviors of teachers and students. The procedures require that observed behaviors be coded or classified by nonevaluative relatively objective sets of categories which describe specific behaviors or actions. Most of the systematic observation instruments in use were originally designed for research purposes rather than for evaluating and measuring teacher performance. Simon and Boyer (1974) in Mirrors of Behavior III report on 99 observation systems. All except one have been used for research purposes. Only twelve instruments are reported to have been used for evaluative purposes.

Many systematic observation instruments have become training tools which give information to people being observed. Simon and Boyer (1974) report that fifty-two of the ninety-nine systems they report on are now being used for training purposes. These systems provide a mirror for the prospective teacher to obtain feedback about his own teaching behavior. This allows the student to attempt to change his own behavior as a result of the feedback. An illustration of this would be interaction analysis. The student would be able to view the extent of interaction in his/her teaching situation. If he/she was dominating the classroom, then the preservice teacher could modify this behavior based on the feedback from the interaction instrument.

A second use of systematic observation instruments in training is that most of the instruments have been constructed along a theoretical dimension which includes behaviors that are thought to promote pupil learning. Many of these behaviors are not found in classrooms usually. This then offers an opportunity for prospective teachers to learn new behaviors and expand their teaching repertoire. Feedback is received on what behavior is not being used as well as on what is being used (Simon and Boyer, 1974).

The implications of using systematic observation to measure teaching competence are many. As has been pointed out, most of these

instruments were developed initially for research purposes. They collect data on specific teaching skills. The data are generally not suitable for evaluative purposes. The instruments are appropriate for use in training programs. They can provide formative data on specific skills that hopefully will help students change their own behavior through this type of feedback.

Some illustrations of systematic observation instruments are the Observation Schedule and Record (OScAR 4V) (Medley, Note 1), Flanders System of Interaction Analysis (FSIA) (Flanders, 1966), and the Galloway System (Galloway, 1968).

Rating Scales

Good (1959) defines "rating" as an estimate according to some systematized procedure of the degree to which an individual person or thing possess some characteristic. He defines "rating scales" as a device used in evaluating products, attitudes, or other characteristics of instructors or learners. Rating scales used for observation are different from systematic observation instruments in that they record general impressions as to what has occurred, whereas the systematic observation instruments describe in detail what has happened.

Rating scales are used more extensively than any other device in attempting to measure teaching competence by observation. Often rating scales include items that are not operationally defined, do not allow for proper time sampling, and do not have interrater reliability.

Remmers (1963) suggests five criteria for judging the appropriateness of rating scales as measuring devices.

1. Objectivity. Use of the instrument should yield verifiable, reproducible data not a function of the peculiar characteristics of the rater.

2. Reliability. It should yield the same values, within the limits of allowable error, under the same set of conditions.

3. Sensitivity. It should yield as fine distinctions as are typically made in communicating about the object of investigation.

4. Validity. Its content, in this case the categories in the rating scale, should be relevant to a defined area of investigation and to some relevant construct.

5. Utility. It should yield relevant information and should not be so cumbersome as to preclude collection of data at a reasonable rate.

Figure 1 is an illustration of an all encompassing rating scale to measure teacher competency. It deals with the intellectual ability of the preservice teacher, includes personality characteristics as well as considering teaching skills. The rater is forced to respond on a five point scale. Figure 2 illustrates a rating scale with descriptors of the competency being measured.

Figure 1
BETHANY COLLEGE, LINDSBORG, KANSAS

Competency Assessment

Name..... Date.....

School..... Subject/Grade.....

Cooperating Teacher..... College Supervisor.....

Consider each of the following items as applying to the student who is seeking certification as a teacher. Mark only one choice for each item by circling the appropriate numeral according to the following scale:

5 - This competency is frequently evident as (s)he functions in the role of teacher.
4 - This competency is sometimes evident as (s)he functions in the role of teacher.
3 - This competency may be possessed by the teacher, but it was not observed.
2 - The lack of this competency is sometimes evident as (s)he functions in the role of teacher.
1 - The lack of this competency is frequently evident as (s)he functions in the role of teacher.

I. INTELLECTUAL

1. Possesses a breadth and depth of knowledge encompassing those aspects of learning for which certification is sought.....5 4 3 2 1

2. Possesses a knowledge of contemporary developments in curriculum materials for the area(s) of certification sought.....5 4 3 2 1

3. Demonstrates ability to determine home and community backgrounds of pupils in order to determine (a) reasons for specific pupil behaviors (b) unique needs of learners.....5 4 3 2 1

4. Demonstrates the ability to devise and carry out systematic self-evaluation.....5 4 3 2 1

5. Possesses a well-developed philosophy of education which he/she can articulate to others.....5 4 3 2 1

II. PERSONALITY

1. Exercises self-control under trying conditions (emotional maturity).....5 4 3 2 1

2. Possesses physical stamina.....5 4 3 2 1

3. Demonstrates in his/her personal life a sensitivity to community expectations in the interest of improved support for education.....5 4 3 2 1

4. Gives evidence of an integrated value system by which life is lived.....5 4 3 2 1

5. Demonstrates integrity in his/her dealings with people....5 4 3 2 1

6. Is willing to expend time and energy as needed in fulfilling teaching responsibilities.....5 4 3 2 1

7. Is punctual in meeting appointments and keeping schedules.....5 4 3 2 1

8. Is able to establish rapport and a professional relationship with parents, pupils, and peers.....5 4 3 2 1

9. Earns rather than demands respect.....5 4 3 2 1

10. Shows respect for the learner.....5 4 3 2 1

11. Demonstrates a sense of humor: the ability to laugh at self.....5 4 3 2 1

III. TEACHING SKILLS

1. Maintains an effective balance of freedom and responsibility in the classroom.....5 4 3 2 1

2. Is able to develop both long range and short range instructional plans.....5 4 3 2 1

3. Is able to use a variety of methods which promote learning.....5 4 3 2 1

4. Is able to develop and use realistic and well-articulated instructional objectives.....5 4 3 2 1

5. Is able to locate and/or devise appropriate learning materials.....5 4 3 2 1

6. Can design and use a variety of techniques in evaluating the attainment of instructional objectives.....5 4 3 2 1

7. Provides for systematic feedback about performance on subject-matter assignments.....5 4 3 2 1

8. Possesses teaching skills to the extent that he/she is able to carry out teaching responsibilities while maintaining a well-rounded life style.....5 4 3 2 1

Figure 1 continued

IV. MAKING KNOWLEDGE MEANINGFUL

1. Is able to identify pupil's level of learning and to prescribe the appropriate learning experiences.....5 4 3 2 1
2. Provides opportunity and guidance for pupils to become independent learners.....5 4 3 2 1
3. Communicates effectively verbally and non-verbally.....5 4 3 2 1
4. Is able to explain, illustrate, and/or clarify tasks to promote efficiency and effectiveness in learning.....5 4 3 2 1
5. Is able to operate and use effectively appropriate hardware and software.....5 4 3 2 1
6. Is able to organize content as necessary to assist learners.5 4 3 2 1

V. CREATING A DESIRE FOR LEARNING

1. Is enthusiastic about his/her subject.....5 4 3 2 1
2. Uses honest and genuine encouragement to increase pupil's self-esteem.....5 4 3 2 1
3. Is an inquirer able to excite other learners.....5 4 3 2 1
4. Has empathy for learners.....5 4 3 2 1
5. Thinks creatively.....5 4 3 2 1

VI. CREATING A DESIRABLE LEARNING CLIMATE

1. Is able to analyze teacher behavior and predict pupil behaviors generated by it.....5 4 3 2 1
2. Is able to determine causes rather than mere symptoms in improving undesirable pupil behavior.....5 4 3 2 1
3. Maintains classroom discipline conducive to learning.....5 4 3 2 1
4. Utilizes good housekeeping habits in care of classroom and equipment.....5 4 3 2 1
5. Organizes and arranges classroom to promote learning.....5 4 3 2 1
6. Uses a variety of motivational techniques in stimulating pupil learning activities.....5 4 3 2 1

Comments, including you recommendation, or lack of it, and the evidence which supports it:

Figure 1 continued

1.101 STATEMENT OF COMPETENCE: Observes pupils in and out of the instructional setting.

FROM				TOWARDS
1	2	3	4	5
Indicators of Competence				
a. Observes pupils in classroom and in other situations.		a. Reports general observations to the supervising teacher.		a. Reports specific observations to the supervising teacher.
b. Uses devices such as checklists to aid in observation.		b. Prepares anecdotal records.		
		c. Prepares written or oral reports of group interaction, individual interaction, and adult-child relationships.		

Other indicators of competence are to be negotiated according to the context in which the student is working.

Figure 2*

* This is adapted from: An Assessment System Utilizing Teacher Competencies for Differentiated Staffing. Developed by Mentor Ohio Exempted Village School District and the Center for Competency Based Education, University of Georgia, 1974.

The rating scale is probably the most useful to teacher educators who are attempting to do both summative and formative evaluation. It is necessary to state competencies in clear and precise terms so that the items to be measured by the rating scale will have operational definitions that leave little room for disagreement. Standard procedures for establishing validity and reliability should be used in the development of any scales.

Interviews

A technique to obtain information that can be used in measurement is the interview. The interview is really a verbal test where preservice students respond to questions related to the competencies they are attempting to demonstrate. A problem with the interview technique is in communication from the interviewer to the respondent and back again. A situation needs to exist where openness and honesty make answers come easily. The place of the interview in measuring competence is in the author's opinion an opportunity to check out other sources of data (e.g., rating scales). One illustration of the interview technique being used to determine competency in beginning teachers is in Dekalb County, Georgia (Note 2). The interview is only part of an evaluation effort with beginning teachers. The interview precedes classroom observations where rating scales are also used to measure teacher competence. Figure 3 illustrates a structured interview instrument.

Interviews cannot serve to evaluate the total competence of a preservice teacher. However, they do provide another source of data that can be part of a measurement effort. It would be necessary to have an interview instrument such as the one displayed in Figure 3 that would provide structure. It would also be necessary to have a procedure for assigning some value to the answers obtained in the interview.

Products

Products, things preservice teachers make such as lesson plans, and anecdotal records, are another source of data that is often overlooked in assessing the competence of teacher education students. The Oregon College of Education's program (Schalock, Kersh, Garrison, 1976) uses ratings on the products of a prospective teacher's behavior as one of five different sources of data in arriving at judgements about competence. Two of their competency clusters rely on the ratings of products. These competency clusters are: Planning and Preparing for Instruction, and Obtaining and Using Pupil Outcome Information. The products that serve as a basis for judgement for the first competency are instructional plans. The products that are used to determine competence for the latter competency are summaries of pupil outcome data, accompanied by interpretations of those data in terms of contributing factors and implications for the next steps. An illustration of a form to measure a product is found in Figure 4. A five point rating scale is used to evaluate the lesson plan presented in that figure.

Products like interviews cannot give a complete picture of a

PBC/SS Project

TEACHER INTERVIEW/QUESTIONNAIRE

COMPETENCY IV - The teacher develops instructional plans for promoting achievement of the specified objectives.

YES	NO	
____	____	1. Do you use varied materials and equipment in your class?
		a. Explain how you organize and use materials and equipment during your class.
		2. Do you use materials for:
____	____	a. remediation
____	____	b. development
____	____	c. enrichment
____	____	3. Do you use evaluation techniques other than written tests or questions? e.g. oral questions, large or small group discussion.
		a. Please describe.

COMPETENCY V - The teacher utilizes resource material.

YES	NO	
____	____	1. Do you have materials for student enrichment or projects?
		a. Describe
____	____	2. Do you encourage students to use the Learning Resources Center?
		a. Explain.

Figure 3*

*Adapted from Teacher Interview/Questionnaire used by the Dekalb County, Georgia Performance Based Cerfification/Supportive Supervision Project.

LESSON PLAN EVALUATION FORM

Student's Name _____ Lesson Number _____

Have the evaluators that check your plan initial each of the items listed that meets with their approval. If the treatment of an item is thought to be outstanding, have the evaluator draw a circle around his or her initials. Be sure to attach this sheet to your lesson plan.

ELEMENTS OF THE PLAN	EVALUATORS OF THE PLAN		
	School Supervisor	College Supervisor	Content Specialist
OBJECTIVES			
Are the learning outcomes expected from the lesson clearly stated?			
Are they appropriate and worthwhile outcomes, given the characteristics of the pupils to be taught?			
Are the indicators that are to be used as evidence of successful outcome achievement identified?			
Are the procedures to be used in obtaining evidence of outcome achievement identified?			
ADAPTING OBJECTIVES TO LEARNER CHARACTERISTICS			
Are there provisions for modifying the objectives of the lesson to meet individual pupil characteristics?			
SELECTING INSTRUCTIONAL MATERIALS AND PROCEDURES			
Are the instructional materials to be used in the lesson clearly identified?			
Are they appropriate to the learners to be taught and the learning outcomes to be achieved?			
Are the organizational and instructional procedures to be used in the lesson clearly identified?			
Are they appropriate to the learners to be taught and the learning outcomes to be achieved?			

Figure 4

ELEMENTS OF THE PLAN	EVALUATORS OF THE PLAN		
	School Supervisor	College Supervisor	Content Specialist
EVALUATION			
Are there provisions for determining where pupils stand with respect to the desired learning outcomes of the lesson before it is presented?			
Are there provisions for feedback to pupils about their performance during the time the lesson is being presented?			
Are there provisions for determining where pupils stand with respect to the desired learning outcomes of the lesson after it is presented?			
PLANNING NEXT STEPS			
Is there some indication in the plan of what would be done next with the pupils if the learning outcomes expected from the lesson materialize?			
Is there some indication in the plan of what would be done next with the pupils if the learning outcomes expected from the lesson did not materialize?			
MATCHING INSTRUCTOR, LESSON AND CONTEXT			
Does the lesson as planned appear to be feasible and appropriate to the school setting in which it is to be presented?			
Does the lesson as planned appear to be feasible and appropriate to the student who is to present it?			

Figure 4 continued
Field Test Format 2
Experimental Elementary Teacher Education Program
Oregon College of Education
December 1972

competency. They do, however, provide a very concrete source of data. The lesson plan, for instance, is something tangible; it can be considered in terms of predetermined specifications and a value assigned. Measuring the value of products requires that a structured form be developed. Usually products that are rated unsatisfactory are returned and not accepted until they are satisfactory. Thus, the rating scale of products is probably formative when ratings are unsatisfactory and summative when they are satisfactory.

Some Tentative Conclusions

The systematic observation systems that have been developed will continue to offer the best means of looking at teaching by the researcher. Some of these systems offer much to the teacher training program. They do not, however, offer a viable means of measuring the competence of preservice teacher education candidates.

At the present time, it would appear that the most appropriate instruments for measuring the competence of preservice teachers are rating scales which can be utilized to make estimates of the value of performance in the work context; structured interviews which assist in verifying what was observed in the work context; and instruments (a type of rating scale) which measures the quality of products developed prior to or during the work context. These data gathering devices used collectively offer a greater probability of an accurate measure of competence than do either of the three individually.

In the development of the three measures care must be taken to define operationally what is to be measured. In the observation rating scales the categories to be used must be specific and leave little to the imagination of the rater. The specifications for the products to be developed must be precise. If a lesson plan is a product, then the elements of the lesson plan must be spelled out so that there is no question as to its appropriateness. The interview instrument must contain questions that relate directly to the competencies being measured.

The context in which the competency is to be measured must be spelled out. If, for example, the preservice teacher is to be observed teaching an inquiry lesson in the structured work context, then the rating scale should be designed to measure this. The rating scale should not be a general scale that attempts to determine how well a preservice teacher does when teaching a lesson. Closely related to the problem of context is what the instrument is designed to measure. The instruments that are developed should be designed to measure specific competencies, again, such as the inquiry lesson rather than a general lesson.

Once instruments are developed, they can be improved only through rigorous field testing. Standard procedures for establishing validity and reliability should be followed. Various time samplings should be undertaken in order to determine what amount of measurement is necessary in order to get an accurate picture of the competence of the preservice teacher. Those that are to be used in the measurement process need to be trained in the use of the instruments. Interrater

reliability needs to be established as well as the amount of time required to use the instruments.

The measurement of competence will continue to be an elusive goal. As was pointed out earlier in the paper, we need to begin. Unsophisticated measures can become sophisticated only if teacher educators are willing to begin and to employ procedures that allow for continuous revision and improvement.

REFERENCE NOTES

1. Medley, D., et. al. Coding Teachers' Verbal Behavior in the Classroom: A Manual for Users of the OScAR 4V. New York: Office of Research and Evaluation, Division of Teacher Education, City University of New York, (No date).

2. Handbook for Field Testing the Performance-Based Certification/Supportive Supervision Model, 1975-1976. Dekalb County, Georgia: (No date).

REFERENCES

Cureton, E. Measurement theory. In R. Evel (Ed.), Encyclopedia of Educational Research (4th ed.). New York: Macmillan, 1969.

Drummond, W., Kennamer, L., Lindsey, M., Massanari, K., Shearron, G. Performance-Based Teacher Education: A 1975 Commentary. PBTE Series, No. 19. Washington, D.C.: American Association of Colleges for Teacher Education, 1975.

Flanders, N. Interaction Analysis in the Classroom: A Manual for Observers. Ann Arbor, Michigan: School of Education, University of Michigan, 1966.

Furst, N., & Hill, R. Classroom observation interaction. In L. Deighton (Ed.), Encyclopedia of Education. New York: Macmillan, 1971.

Galloway, C. A Description of Teacher Behavior: Verbal and Nonverbal. Columbus, Ohio: Ohio State University, 1968.

Good, C. (Ed.). Dictionary of Education (2nd ed.). New York: McGraw-Hill, 1959.

Johnson, C., Shearron, G., & Hensel, N. A Meaning for Competency. Athens, Georgia: Competency-Based Education Center, University of Georgia, June, 1974.

Kay, P.M. Measurement techniques: What we have and what we need. In R.W. Houston (Ed.), Exploring Competency-Based Education. Berkeley, Calif.: McCutchan, 1974.

McNeil, J.D., & Popham, W. The assessment of teacher competence. In N. Gage (Ed.), Second Handbook of Research on Teaching. Chicago: Rand McNally, 1973.

Remmers, H. Rating methods in research on teaching. In N. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Schalock, D., Kersh, B., & Garrison, J. From Commitment to Practice: The Oregon College of Education Elementary Teacher Education Program. PBTE Series, No. 20. Washington, D.C.: American Association of Colleges for Teacher Education, 1976.

Shearron, G., & Johnson, C. A CBTE program in action: University of Georgia. Journal of Teacher Education, Fall 1973, 24, 187-193.

Simon, A., & Boyer, E. (Eds.). Mirrors for Behavior III: An Anthology of Observation Instruments. Philadelphia: Communications Materials Center, Research for Better Schools, 1974.

Turner, D. Evaluating the validity of assessed performances: Methodological problems. In R.W. Houston (Ed.), Exploring Competency-Based Education. Berkeley, Calif: McCutchan, 1974.

CONFERENCE SESSION

Gilbert F. Shearron
University of Georgia

Question: What is the problem, what is the danger, in being overly specific or too specific in the development of rating scales?

I think there are a number of dangers in being too specific. Several years ago we became concerned about this business. A lot of people had the idea that the only way you could write evaluation items was to write objectives using the Mager framework; you could only use certain kinds of adjectives, certain kinds of verbs, etc., and then you had a behavioral objective. The more I got involved in that, the more I found that we came up with things that were not very important. You get down to so many little bits and pieces that it is extremely difficult to put them together and come up with a whole that really measures something.

I think the major danger is that we often break a competency down into so many parts that we forget about the whole. I think that is one of the reasons I made those comments in the paper about the use of systematic observation instruments. They tend to measure bits and pieces in a highly specific manner as opposed to what I'd call a total kind of competence.

Question: Would we modify the teacher interview questionnaire that appears on page 14 for use with special education people and, if so, how?

Yes. I think there would probably be a number of additional questions that should be asked for special education teachers. I think you probably would want to talk about the level that the materials were appropriate for; you'd probably want to talk about how these materials might be used in individualization. You could include a number of such questions.

Question: Would you comment on the use of the interview scale to collect data?

If I were going to use this interview scale, I would write a narrative as the student responded. If you only use Yes and No you would probably not get the kind of data that you would want to get.

I believe very strongly, and I hope my conclusions in the paper bring this out, that there have to be multiple sources of data to

really do measurement as it should be done. I think the interview scale certainly can be a backup for a rating scale, but I also think that an evaluation of products can be a backup.

Question: Suppose the interview and the rating scale provide different kinds of data; how do you resolve that?

I think the only thing you can do is go back to your interview scale and try to clarify where these differences were with the person who is being evaluated. For example, if the student says that they are using, oh, let's say, a variety of materials in teaching a particular lesson, and then you go in and find out that they are only using one procedure and everybody is doing the same thing, then obviously the student either doesn't know what you are talking about, or, is trying to mislead you. I think we have to give the student credit the first time around--maybe even the second or third time--by going back and asking him, "You said such and such about the use of a variety of materials; now we didn't see you do that. Perhaps we need to talk about what you mean by a variety of materials, and what we mean." Again, it gets back to saying things in some operational framework so that it does communicate between the person being evaluated and the person doing the evaluation. I believe the interview can be used in a formative sense, as well as the other instruments. Again, I want to say that I guess I'm committed to what I consider to be the need to have multiple sources of data before making a final decision on competence.

Question: Do you believe that high inference measures should be used for summative evaluation and low inference measures for formative evaluation?

Low inference measures may be more reliable or valid than high inference measures--from a researcher's point of view. I personally feel that the state of the art right now is that we use a lot of high inference stuff, whether it be for formative or summative evaluation. I think we are making a lot of subjective judgments. I don't have any problem with that personally because I think that's where we are. I think in teacher education programs for the past 50 years, or since we've had teacher education programs, we have used various subjective kinds of judgments about whether people could teach or whether they could not teach when they finished their undergraduate programs. I personally think there is room for both high inference and low inference in both the summative and the formative evaluation, but at this point I think most of the measures that we use would be high inference. The kinds of things I'm talking about in the paper are basically high inference measures, and I don't make any apology for that; I think that's where we are.

Question: Do we assign some sort of numerical value to a lesson plan evaluation form so that we can make some kind of judgment?

Yes, you probably do need to define some numerical value to the elements of the lesson plan, if you think that is important. That

is the kind of a local decision that needs to be made. I feel very strongly that CBTE is a process. I think you can move the process around from institution to institution, but I think decisions like this have to be made at the local level, again depending upon what you feel is important.

Question: Should everybody be expected to master everything that is on the evaluation form, or should there be some magic percentage?

This is another question that I think has to be dealt with at the local level. I don't think that you can expect 100% mastery; that's just a personal opinion.

Question: How would I respond in a case of due process by a student when using the data sources you recommended?

I'm responding now to your question in terms of a student in an undergraduate teacher education program. In our own program, we gather data from three sources: peer evaluation, university staff (two university people, one of whom I would call a generalist and one I would call a specialist), and the public school teacher. After the evaluation takes place, the student has the opportunity to sit down with the people who have done the evaluation and discuss what has happened.

We try to record what was said in that particular conference. If the student is not satisfied, we go back and there is another evaluation. If the student still is not satisfied, then that student has a right to present his or her grievances to a grievance committee. If the committee upholds the people who did the evaluating, and the student wants to go to court about this (and we have not had one do this), what we propose to do is to say, "Okay." We went through this process. We had data from three sources. I don't know how it would be adjudicated. If you look at the court cases, it would probably have a 50-50 chance of going either way. The best thing to do, if you think the student has not done what he ought to, is to give him a low grade. If you put a grade on there, an F for example, I think you are safer than if you just talk about some sort of data where you say this person was satisfactory or unsatisfactory.

Question: Were you suggesting then that other grades be used as opposed to just S or U?

I'm suggesting that in the summative evaluation you do have to assign grades, letter grades, yes. Conceptually, I don't think that's a good idea, but practically, I believe we are on safer ground with the letter grade.

Question: Have you had experience with how many times you let a student attempt to achieve a competency as whatever level?

I have to answer that by my own experience. We try to do extensive testing of all of our students before they are admitted to the program.

I mean a whole battery of tests where we collect data and so forth, and then they go out in the field and spend about a quarter deciding whether or not they think they like teaching. We have a number of them who opted out after they saw the kind of situation they'd be in. One of the things we have our students do is to try to do a lot of diagnostic work. We have let them work with up to five different youngsters trying to see if they could demonstrate that particular skill. I don't recall competency demonstration ever going over five. If after five attempts they still can't do it, we usually try to counsel them into something else. That is basically what we do in the structured context of our program. In the unstructured context, we have extended the time that a student would be out in the field where they might have a number of opportunities to do this. This is always dependent upon the school situation that you find yourself in. In that unstructured context, sometimes the school people are not really open to a student constantly making the same mistakes. I think we give most of them a second shot, some a third. I really don't want to go beyond that. But it depends upon the situation; it depends upon whether or not you see improvement the second time around. I don't think there is any magic way that you can decide that.

A REPLY TO DR. SHEARRON'S PAPER

James D. Strauch
University of Connecticut

Professor Shearron discusses a wide range of issues concerning the development and improvement of instruments for measuring competence in CBTE programs. His assignment was complex and could cater to only some of the most prominent issues in measuring the behavior of preservice teacher education students.

He notes that "...we are concerned about the measurement of performance of people and the assignment of some sort of value to that performance." He also states that "performance is an observable behavior, but it includes the manipulation of ideas and the making of judgments and decisions." In his presentation, performances are subsets of teaching competencies, and he emphasizes that the major contexts for measuring teacher behavior should occur in a simulated or work context a la Turner.

Before proceeding into an examination of actual instruments, two issues should be highlighted. First, we are assuming that in one manner or another a set of competencies has been identified for the special education teacher training program. There are numerous commonalities between general and special education teachers, but there are also differences. It may be demonstrated for example, that the special education trainee must be more competent than the general educator in the areas of diagnosis and prescribing, arranging instruction and/or the classroom environment for disturbed children. Working with groups of children with multiple handicaps would be another presumed competency for some special education trainees.

Special educators are in the process of identifying sets of teaching competencies that are unique to working with the handicapped children. (Note CEC, Competency Based Teacher Education and Evaluation, 1975). Only when competencies are identified, even at the level of face validity, will special educators be able to proceed in designing appropriate measures of performance. Existing measures, whether they be observational schedules, rating forms, or other devices are appropriate to the extent that they correspond to the competencies of a particular program. Assuming training programs in special education are in some ways different from other programs, it is essential that behaviors being taught in training programs be incorporated into the measurement devices. Many of the currently available observation schedules, rating forms and other measures of teacher performance do not adequately interface with sets of various special education teacher competencies that have been identified.

A second issue that still needs resolution, and raised by Professor Shearron, concerns the level of specificity that competencies should be stated in. Considerable information has already been published by the Multi-State Consortium on Performance-Based Teacher Education and the Amer-

ican Association of Colleges for Teacher Education, in addition to numerous texts and journal articles on how to design CBTE programs. What is still lacking, however, is a taxonomy or hierarchy for organizing competency statements that will allow for equivalence across and within programs. In other words, when is an intermediate level competency an intermediate level competency? Is a competency statement in one program equivalent (e.g., degree of specificity) to that in another program? The organization of such a taxonomy or hierarchy of teaching competencies would assist teacher trainers in designing manageable evaluation schemes and reduce the number of false starts and redundancy of effortin program evaluation. The need for such a hierarchy has been discussed more fully elsewhere (Strauch, 1974).

For all practical purposes, the problems faced by educators in evaluating their trainees (and trianing programs) are comparable to the myriad issues in evaluating any cirriculum.

Measuring knowledge, attitudes and other symbolic processes at the college level is familiar to the readers of this paper, and will not be dealt with here. It is more important to dwell upon ways to evaluate teaching competence in simulated and work contexts--structured and unstructured, as discussed by Professor Shearron. Again, special educators will especially want to attend to the work contexts in which evaluation is to occur--special class, resource room, residential facility, etc.

Since the majority of training programs are still in the process of development, and given constraints of time and resources, it is incumbent upon program planners to sample a trainee's performance in whatever work contexts are deemed most critical, both simulated and the real class. It is unrealistic to expect emerging programs to measure the performance of all trainees relative to all program competencies. As Shearron and many others have pointed out, measurement development concerning teacher performance is still in its infancy (Gage, 1974; Heath & Nielson, 1974; Houston, 1974; Kay, 1974). At best, teacher education programs in special education will do well to focus upon criterion-referenced means of assessment which, in many cases, means tailor-made measures for particular programs. There are very few ready-made instruments available that will be sufficiently prescriptive to assist in changing trainee performance. And while the unit of sampling may be the child being instructed, in most instances we would do well to first focus upon sampling the performance of our trainee. Our training endeavors are not yet at a point of demonstrating unequivocal effectiveness using pupil performance as a criterion measure. The job of the moment is to evaluate in valid and reliable ways whether our trainees possess the competencies that make up particularized training programs.

The criterion-referenced approach will require the setting of priorities among stated competencies, sampling of performances of trainees using the best measurement techniques available.

Particular techniques--rating, observation, interview, and others should contain specificity as Professor Shearron suggests. The instruments used for data collection should be low inference, especially as the training program is being developed. Low inference measures can provide accuracy to both the teacher educator and trainee since the items are, hopefully, clear, discrete, and probably few in number. Basically, the low inference

measures would be utilized in process evaluation.

The high inference measures, including the displays in Figure 1 and 4 of Shearron's paper can be used in summative types of evaluation. These measures are more gross, provide a greater capture of behavior but are less reliable than the low inference measures, and are less useful in prescribing changes in a trainee's performance.

General and special educators have yet to determine what teacher behaviors will promote particular pupil outcomes. This is an ultimate goal for all educators. In the meantime, CBTE programs should be able to identify competencies, prioritize them, select contexts in which trainee performance will be measured, and sample sets of performances using both low and high inference measures.

In the final analysis, professors and supervisory personnel must make value judgments about the performance of a trainee. The measurement techniques alone, especially at this stage of their development, are incomplete. As a first step we should make a strong effort to demonstrate our competence as teacher trainers by measuring program effectiveness through trainee performance.

REFERENCES

Competency-Based Teacher Education and Evaluation: A Selective Bibliography. Reston, VA: The Council for Exceptional Children, 1975.

Gage, N.L. (Ed.). Mandated Evaluation of Educators: A Conference on California's Stull Act. Washington, D.C.: Education Resources Division, Capitol Publications, Inc., 1973.

Heath, R.W., & Nielson, M.A. The research basis for performance-based teacher education. Review of Educational Research, Fall 1974, 44(4), 463-484.

Houston, R.W. (Ed.). Competency Assessment, Research and Evaluation. Syracuse, NY: Syracuse University, 1975.

Kay, P.M. Assessment in CBTE: The Search for Evidence. New York, NY: Graduate School and University Center of CUNY, 1974.

Strauch, J.D. Delineating, training and evaluating particular competencies in special education. In J.J. Creamer and J.T. Gilmore (Eds.), Design for Competence-Based Education in Special Education. Syracuse, NY: Syracuse University, 1974.

TECHNIQUES AND CRITERIA FOR DESIGNING AND SELECTING INSTRUMENTS FOR ASSESSING TEACHERS*

Paul Pottinger
Institute for Competence Assessment
McBer and Company

The Problem in Perspective

More than ever, teacher educators want to know and need to demonstrate whether or not they are accomplishing the goal of preparing people effectively for teaching. The development and use of assessment and evaluation techniques, however, have not kept pace with the need for better answers to these fundamental questions.

Three years ago, David Krathwohl (Merwin, 1973) stated:

> One can predict that performance-based teacher education (PBTE) is certain to fail to reach its ultimate objective if it continues on its present course. This failure will be caused by the almost complete lack of attention given to the assessment of teaching competencies, a core concept of PBTE. Only by such assessment can we achieve the goal of assuring that a teacher can indeed perform in ways that result in children learning.

Merwin went on to say in this same publication:

> Indeed, the ability of trainers to adequately define and measure competency lies at the heart of the question of whether PBTE can be successful.

Dr. Krathwohl further commented that there is a paucity of instruments available for assessing teachers. This is still true today, and it is one reason, no doubt, that I have been asked to address this group on criteria for development and selection of instruments. My comments will focus primarily on development, partially because I am aware of few relevant instruments from which to select.

Others have written eloquently about the problems of PBTE (see especially Broudy, 1972 and Merrow, 1974). Still others have discussed issues of assessment research which provide useful criteria in the development or selection of instruments (see for example, Merwin, 1973; and Medley, Soar and Soar, 1975). I will not attempt to summarize these readings since you are no doubt familiar with them and many others already.

* Originally appearing in B. Levitov (Ed.) Licensing and Accreditation in Education: The Law and State Interest. Lincoln, Nebraska: University of Nebraska Printing and Duplicating Service, 1976. Reprinted here by permission of the author.

This presentation will:

-- briefly discuss a variety of perspectives that impinge upon assessment and performance-based teacher education.

-- suggest a conceptual model for meeting the simultaneous demands of those who have different perspectives yet are in search of a common ground for achieving practical solutions for validating teacher competencies.

-- discuss some specific assessment instruments, developed or utilized in a variety of educational and occupational settings, which serve as prototypes and models for new instrument development and selection.

Background-Social vs Scientific Impetuses for Change

The impetus to change assessment of teacher performance has come primarily from the courts, taxpayers, and educational consumers. This impetus has culminated, in New York State at least, with Commissioner Nyquist's challenge that we must assure the public that graduates of accredited teacher education institutions are more competent than non-graduates. Performance-based teacher educators are trying to instill public confidence in teachers' competence by empirically validating specific behavioral objectives. Research, theory, and development, however, have not kept pace with the social, political, and legal requirements for change that are being imposed.

While we should welcome incentives for change, we should recognize them as a mixed blessing. New York's validation goal is a frightening proposition for higher education researchers who find little support for the rigorous research needed to meaningfully validate teacher competencies. Government and private sector funding sources as well as program developers and researchers are responding more to the social forces for change than to rigorous empirical (i.e. scientific) quests for better, more practical solutions to complex problems. PBTE efforts have reflected a reactive movement in response to simple and minimal court demands for content validity and quick answers rather than a proactive movement in response to fundamental scientific demands for better knowledge of cause and effect relationships. If we attempted to answer these more fundamental questions, the issue of predictive validity, rather than mere content validity, would receive more attention than it now does.

Thus, social pressures have led educators into attempting to clarify objectives, goals, and outcomes without first establishing a strong methodological base from which to make reasoned decisions. Such decisions are instead being made from considerations which lack the vision, concepts, and methods necessary to meet meaningful, long-range assessment goals.

Performance-based teacher education appears to be a rational approach for making program and individual certification decisions. Its efforts to date, however, have fallen quite short of meeting the needs for a system that is not only defensible in the courts but, perhaps more importantly, defensible to students, teachers, taxpayers, program developers, and others who want to know what constitutes effective teaching.

The courts are not requesting the quality research needed to answer fundamental questions of cause and effect relationships because their analyses, albeit systematic and rigorous, are aimed instead at assuring due process. That is, the court's rightful role is to demand fair processes of decision-making within a framework of what is known. The researcher's role, in contrast, is to improve upon this knowledge base rather than to assure fair and equitable use of it. As long as educators respond to the rigor and logic of due process rather than to the rigor and logic of science, PBTE will not withstand the challenge of the courts any better than standard or traditional educational processes.

One cannot ignore the social pressures for change which arise from legal, economic, political, and other social conditions of the day. But we must be clear about the differences in goals and strategies for change which these forces create. The education researcher may be interested in validity, theory, methodology and other scientific concerns which are of little interest per se to taxpayers or the courts. A teacher may be primarily interested in being fairly assessed for a broad range of knowledge, skills, abilities and other characteristics that constitutes his/her competence; the teacher is appropriately more interested in comprehensive and fair assessment than in scientific issues. The education administrator must be concerned about accountability demands from accrediting organizations. The State Board of Regents is concerned with accountability demands from the courts and the public in general. And the student and his parents want more for their consumer dollar than is often demonstrated by our teaching institutions. Each perspective is legitimate but no one perspective should be allowed to dominate our strategies for solving the problems which ultimatey each of us shares. Adequate criteria for selection and development of assessment instruments should keep these differing and relevant perspectives in balance. To maintain this balance, we need competent integrators who can translate the critical needs of practitioners, administrators, parents, students, and the government into practical scientific approaches to problem solving.

The Researcher's Perspective

For the researcher, a more salient problem than maintaining a balanced perspective is, frankly, defining what competence is. And that is a particularly delicate issue if one considers measuring teacher competence in terms of student outcomes. Let's look briefly at some real-world barriers to answering this question satisfactorily.

The objective of schooling, for example, becomes relevant to defining competence in terms of student outcomes. Is the objective to foster the general transition of children into adults? Is it to prepare people for the world of work? Is it to function as a socio-political-economic gatekeeper with respect to who attains further educational and career opportunities? These questions represent philosophical and ideological strata which are more properly the domain of values than of science.

Assuming there is agreement about the objectives of schooling, we would then ask what student outcomes are appropriate to these objectives. Are they primarily having a grasp of the three R's? Or do they include interpersonal competencies? Are they critical thinking and problem solving

abilities? Are they moral behaviors? Do they include the constructive and effective use of power or an orientation toward achievement? If we can answer the first set of value-laden questions, we can at least begin to answer the question of appropriate outcomes with some empirical rigor. Until we have done so, we cannot very meaningfully ask what teacher behaviors and other characteristics best enable these learning/development outcomes. Yet, little progress has been made toward developing rigorous, useful and relevant operational definitions of student outcomes. No wonder there is simply no good evidence to relate teacher behaviors to student outcomes.

The reason there is so little evidence to provide us with fundamental knowledge about what works is, of course, the sheer complexity of the problem. There are too many unknown or uncontrollable variables. Students differ in their basic behavioral repertoires, socio-economic back grounds, motivations, interests, values, and perceptions of their environments. Teachers also differ according to these and other important dimensions. Finally, the environments within which teachers and students interact differ according to numerous physical, psychological, organizational, social, economic, and political dimensions. Medley, Soar and Soar (1975) have eloquently presented a model for conceptualizing the complexities created by these variables, and they have elaborated on the implications of these obstacles for doing research. We will refer to their four-level paradigm for conceptualizing research in the next section.

One reason for elaborating on the socio-political context of competence assessment as well as on the scientific complexity of the problems is to make two points more salient. That is, (1) we cannot meaningfully discuss criteria for developing or selecting instruments for assessment without agreeing upon the context or paradigm from which we operate together. Part II of this presentation will briefly describe the paradigm from which I am operating. Suffice it to say here that selecting and developing measures of competence is to my way of thinking, putting the cart before the horse, since there is little or no scientific evidence about what constitutes competence in the first place. (2) More systematic collection and analysis of data is required of all of us if we are to make significant progress toward the goal of empirically validating teaching competencies by 1990. We need not, by the way, agree upon philosophical and ideological issues in order to do empirical research. Also, heavy emphasis by researchers on rigorous scientific analyses of competence should not denigrate practitioners. Many practitioners have strong convictions about what constitutes quality education, but they are unable to empirically validate these convictions. We should not belittle them for assessing student competence on a subjective basis; rather, we should tap their intuitions and covert working hypotheses to uncover clues to the causes of successful teaching.

The plea for more scientific research stems from the belief that such research is critical to the development of quality PBTE programs that are attempting large-scale change in the way they teach, assess and credential students. Moreover, the outcomes of quality assessment research could replace legal and political pressures as the "prime mover" in accomplishing the changes.

Performance-based teacher education will not progress very far unless it is able to move beyond the subjective assessment of the "laundry lists"

of behaviors which proliferate the field. No matter how strongly current curriculum and assessment innovations are supported by the courts and others who demand equity and accountability, PBTE must provide scientific evidence that it works better than the status quo. The barriers to defining the measuring of competence will not be eliminated by arguments for change and innovation no matter how strongly they side with equity, accountability or other relevant social goals. The outcomes of PBTE must be developed and evaluated systematically, rigorously and scientifically.

Most new attempts to define and measure learning outcomes according to what people can do are restricted in scope, lack rigor or poorly correlate with job and life requirements. The state of the art of assessment is in need of more conceptual rigor, systematic and comprehensive strategies for developing measures, and empirical verification of the utility of these measures.

Until we have a more comprehensive base, educators will continue to use an existing array of questionable measures based on narrow cognitive outcomes, superficial behavioral observations, and a priori value-laden judgments. What is required is a sophisticated technology capable of uncovering critical factors which are necessary and sufficient for competent performance.

Criteria for Assessment Instrument Development: Locating a "Rational Center"

Moving from Behaviors to their Causes

The one criterion I wish to emphasize is that teacher competency definitions and assessment instruments should comprise causally related, criterion referenced and generalizable variables which take into account complex observable and unobservable behavioral repertoires required for successful performances. Underlying this statement is intuitive, theoretical, and empirical evidence supporting the notion that competencies cannot be meaningfully defined by seemingly endless lists of specific skills, tasks and actions which ultimately fall short of real-world requirements for effective performance. In fact, the more essential characteristics for success often turn out to be broad or generalized abilities and characteristics; and they are sometimes more easily operationally defined and measured than an array of specific "subskills" which do not add up to general competence.

Most PBTE assessment techniques are based on the observation of external behaviors which are the building blocks of successful performances. But these assessment techniques tend to be reductionistic and lacking in meaning because they fail to assess the underlying causes of these behaviors. Usually the result is the assessment of behaviors which may have little generalizability or transferability to a variety of teaching requirements. This problem has important implications for curriculum and instruction, because observable but superficial behaviors rather than their causal underlying factors often are taught. Thus, what is actually learned, as well as what is assessed, may have little general significance in post-academic life.

This bias will arouse strong feelings in many supporters of the PBTE

behavioral objectives approach to competency definition and measurement, so I will try to put this criterion into an historical as well as methodological perspective.

Again, differing perspectives of program developers must be taken into account. I hope to find some practical middle ground, not because it is safe, but because I think that recognizing the major thesis, the need to measure causal variables, can potentially lead to pragmatic solutions for finding what Merrow (1974) calls a "rational center" in PBTE.

Further Background: Reliability and Specificity vs. Holism and Relevance

A basic tenet of PBTE, as I understand it, comes from the Committee on PBTE in its sixteenth publication entitled, "Achieving the Potential of Performance-Based Teacher Education: Recommendations," (1974). This document states,

> The unique strength of PBTE is that it challenges all who touch it to be open about their intentions and explicit about how they will decide if their hopes are fulfilled. Ends must be made explicit; means must stand the test of relevance. The logic of the performance-based approach places a healthy stress on the use of evidence to test one's ideas and assumptions. In these considerations lie what the Committee believes to be its great potentialities (p.29).

The interpretation and implementation of the spirit of this statement has often come from an accountability (social-economic-political) or due process (legalistic) frame of reference. These interpretations have been partially responsible for the formulation of competence definitions in terms of a plethora of professional judgments, documented by a morass of highly specific but unvalidated competency statements. The lack of rigorously demonstrated relationships between competency statements (or measures) and educational outcomes has not been a serious problem for those who view PBTE as a short term pragmatic remedy to problems of accountability. Nor has this troubled those who view PBTE as a remedy to the courts' request for content valid assessment devices for licensure.

Some program developers have been led away from accounting for the richness and complexity of human behavior in their curriculum and assessment procedures by behaviorism or the "tyranny of reliability." They have fallen prey to the specific behavioral objectives approach because they believe that molecularizing global behaviors into sublists is theoretically sound. Or they believe that this reductionism makes behavior more amenable to objective observation, which, ipso facto, solves the problem of reliability. Merwin (1973), however, pointed out three years ago that one need look only briefly at the most comprehensive collection of statements of teacher competencies available at that time, to ascertain their insufficient implications or directions for objective observation. He concluded that,

> Acting only on these competency statements, it is highly unlikely that two trainers would independently structure the same assessment procedures or that two observers would attend to the same aspects of performance (p.12).

These comments appear to be as applicable to current competency statements as they were three years ago. That is, competency statements are not providing an adequate base for designing behaviorally specific assessment instruments.

Perhaps this failure is a blessing in disguise, for it provides us with incentives to take stock of this atomistic/reductionistic approach. Even if reliability can be increased, which some research studies have found to be the case, the relevance of rated behaviors seems dubious.

Implicit (and sometimes explicit) in this approach is the assumption that the acquisition of microskills and abilities adds up in some linear and simplistic fashion to overall competence. That is, highly valued generalizable skills are assumed to be mere summations of the ability to perform objectively measureable and highly specific sub-tasks.

Broudy's (1972) criticism of this assumption is useful here. The belief that the whole is merely the sum of the parts

> is a notoriously inadequate description of any human action, let alone one so complex as teaching. Teaching can, of course, be thought of as broken down into parts, but as a concrete action it is guided at every moment by a sense of its total pattern. This pattern - in swimming, reading, classifying, judging - integrates the analyzed constituents into a meaningful functional sequence, not merely a mechanically additive one. We are told, at least by some psychologists, that after the pattern has been sensed or felt or understood, the details can be perfected separately, but until the pattern has been discerned, drilling on the separate parts yields disappointing results (p.3).

I think Broudy has written eloquently on this concept of the whole vs. the parts. I shall take the liberty to quote him further. He discusses the limitations of behaviorally specific prescriptions for teaching performance in the context of three teaching styles: the didactic, heuristic, and philetic.

> Didactics refers to the impartation of knowledge by the teacher to the pupil; heuristics refers to the effort to help the pupil discover for himself either the contents of a body of knowledge or the methods of arriving at such knowledge and assessing it; philetics is merely a Greek name for love or securing rapport with pupils or, as the current jargon has it, "relating to pupils."
>
> Performance-based programs can accommodate didactics which aims at more or less rote mastery of a repertoire of explicitly formulated knowledge and skill. Heuristic and philetic teaching do not lend themselves to the precise analysis, specification, and evaluation which is the presumed glory of the PBTE. Apropos of which, one might remind the namers of teaching machines that Plato and Socrates were exemplars of heuristics, not didactics.

When a fairly reliable measure of learning is available - as it is in didactics - we can take a Skinnerian position and say, "Given teacher performance P, there will ensue pupil performance S," and we can perhaps ignore (for heaven alone knows what concomitant learnings take place) whatever intervenes between P and S in the minds and hearts of pupils, teachers, parents, and school boards. This is the tough line adopted by the proponents of behavioral objectives, education contractors and contractees, and the directors of the budget, local, state, and national. Such toughness makes no sense in heuristic and philetic teaching, where learnings are insights and transformations of attitude for which unambiguaous behavioral indices are hard to find, inasmuch as tolerance of ambiguity and lack of structure is an avowed outcome of philetics. What behavior, for example, shall we regard as criterial for a pupil's insight into his hostility to the teacher?

Success in heuristic and philetic teaching cannot be judged by prespecified appropriate pupil behavior because such behavior - even when we can identify it - it is not manifested on demand or at a specific time. Critical thinking, the use of the imagination, warm feeling toward peers, achievement of identity cannot be inferred from one segment of behavior used as a test pattern. And what pattern shall we use as a test? Indeed, the vulnerability of general education to attack lies in the very fact that many of its benefits do not appear until fairly late in life. Our speech and reading habits, a thousand attitudes, our interests often represent the tacit functioning of explicit learning inputs made during school and college, but which we can no longer recall. This may help to explain why correlations between academic achievement and success in life are so low. The academic grades measured learning of items that have since been largely forgotten; functioning now are the residual conceptual and affective schemata, which were never tested on examinations. Nor need it be added that the life outcomes we claim for heuristic and philetic teaching are from the first contaminated by noninstructional variables, which we are never able to control adequately in our research or schooling (pp.5-6).

A conclusion reached by Broudy in this same document is that how often one has performed a specified task or over what range may not be so important as the ability to perform a variation of the task not previously practiced. Broudy discusses this in the context of teacher training, but its implications for assessment strategies are congruent with my central thesis. Unobservable but measurable, generic and causal linking variables of behavior are equally if not more critical than observable behaviors for defining and measuring competence. Broudy's example of a teacher whose task is to explain Boyle's law further captures the essence of this thesis:

Suppose the prospective teacher recited the explanation of Boyle's law verbatim as it was put down in his textbook or the teacher's manual. Suppose he got all his pupils to do likewise. Would not this be proof of performance competence? Suppose, in addition, he could do all the exercises dealing with Boyle's law at the end of the chapter, and suppose most or all of his pupils

could do likewise. What more definite and objective evidence of competence could one want - if that is the competence one wants? Yet it is clear that such a performance could be brought off without either the teacher or the pupils "understanding" Boyle's law. (Indeed, many generations learned geometry in precisely this way.) As a matter of fact, a demonstration that would really satisfy us that "explaining" Boyle's law had been performed adequately would not be any specific prescheduled behavior. On the contrary, some sort of dialogue with pupils that allowed us to infer - not observe- the basic net of concepts we call chemistry is understood by both teacher and pupils is needed. The kinds of examples and counter examples; the way pupil questions are interpreted; the cues used to set the pupil on a more profitable course; not the performance but the state of mind we call understanding is the crucial "product" here. No single observable behavior is likely to be sufficient proof of such adequacy, for a state of mind is not observable behavior. Skinner quite rightly doesn't worry about whether his pigeons understand what they are doing so long as they do it. If however, the way a situation is perceived or interpreted is in any way an important ingredient of teaching or learning, then verbal behavior, or any other covert behavior, may not be sufficient indicators of either successful teaching or learning. In other words, performance-based teaching is in danger of capturing everything except what is most significant in many kinds of learning, viz., significance (pp. 10-11).

The implications of Broudy's comments for developing assessment instruments are many. The basic conclusions might be summarized as follows: The preoccupation with specificity, clarity and precision of behavioral objectives has left us with an operational paradigm for defining and measuring teaching competencies which is intuitively and theoretically oversimplified and invalid. Lengthy checklists of highly specific "low-inference" behavioral outcomes - even when properly sampled, factor analyzed, and interpreted have been found empirically (though not yet legally) insufficient for validating competence. Rigorous behavioral observations fail to adequately assess either complex interactions among variables or underlying causes of competent teaching performances. To often qualities of the mind and character, unamendable to direct observational techniques, account for the presence or absence of competent performance.

The alternative to the behavioral objectives paradigm appears to be an operational system that leads to vague, abstract, and unmeasurable outcomes. Yet, our intuitions as well as empirical evidence tells us that competence is comprised of non-additive, non-linear, complexly interacting variables - some of which are not directly observable (see Medley, Soar and Soar, 1975; and Gage, 1976).

Causally-Related Criterion Referenced Measures: The Rational Center

What is needed, then, is a model for defining, operationalizing, and measuring competence that combines the methodological rigor of the behavioral objectives approach with the richness of our intuitive, empirical, and theoretical knowledge of what constitutes competence.

One model for doing this has been developed by David McClelland, of Harvard University, and his colleagues. Its success has been demonstrated

in a variety of occupational areas. As it is applied to teacher competencies, it assumes that more important than either the storage and retrieval of information or the frequency and range of tasks performed (as specified in behavioral objectives) are hosts of covert behaviors which are antecedent or causally related to successful behavior. For example: How able are people in processing new information for problem solving; how able are they in integrating this information to form new solutions; and how able are they in implementing these solutions.

The distinguishing characteristic of this model is that it attempts to define competence in terms of causal variables of competent behavior rather than in terms of external behaviors that are the observable manifestations or "building blocks" of competence. Such behaviors may correlate with success, but correlation or mere relatedness does not imply causation, a fact almost completely overlooked in the behavioral objectives model of PBTE. Specific behaviors may be statistically associated with teaching success, but it does not follow that these behaviors lead to success; they may have followed it. Thus, ICA's model is aimed toward defining and measuring characteristics which are not only statistically related to outcomes, but in fact, cause them.

A major assumption of this approach is that knowledge and skills that can be defined in terms of behavioral objectives are seldom sufficient indicators of how well a person will perform as a teacher. There are many other factors that relate to performance but are not tapped by behavioral observations. These include, for example, motivation, diagnostic listening abilities, empathy, the ability to think clearly under stress, the ability to anticipate, analyze and solve problems, the ability to be flexible in teaching style, and many others. Often these factors are intuitively obvious as critical to teaching success, but they are rarely measured effectively if at all. It is these and other variables related to complex teacher abilities that causally-related criterion measures are designed to assess.

PBTE has successfully broken away from the limited assumptions that knowledge of teaching theory alone is sufficient for developing competent teachers. But it may have swung too far to the opposite extreme of only teaching behavioral tasks. Neither the knowledge of how one should teach nor simple demonstrations of behaviors gets to the heart of competence. We must establish links between the knowledge and behaviors that constitute successful performances and the causes of these performances.

The ICA Model

What causes a teacher to utilize knowledge and behaviors effectively? ICA's operational model for answering this question would encourage one to raise new hypotheses about these causes and to validate them. For example, if what Broudy calls "insights and transformation of attitudes" or "residual and affective schemata" are important to effective utilization of behaviors, we must try to operationalize and objectively measure these causal variables. Such hypotheses might well come from observing or analyzing the heuristic and philetic aspects of teaching. We must recognize that "sensed patterns of behavior," and "perceptions and interpretations of situation," are causally related to successful performances. Educational researchers must begin to operationalize their hypotheses about these causes and see if their measures differentiate superior performers from average or poor performers. These measures should be objective, reliable, and amenable to statistical analyses,

so they can be construct validated and empirically related to successful performances. (For further discussion of determining the meaning of measures through construct validations and criterion referencing, see Appendix B.)

Attention must be given measurement of what many researchers treat as "extraneous factors." Often it is these factors, considered by some to be merely undesirable "noise" in the system, that contain rich clues to the causes of competent performances. These variables are intuitively recognized as critical aspects of success, but they have too often been treated as extraneous factors, because people don't know how to measure them. Researchers must come to grips with these critical causal components of competence. Otherwise, PBTE will fail to be meaningful to teachers and others who know that more than their repertoires of observable external behaviors are functioning in their effective teaching interventions.

If an assessment instrument validly taps a causal variable (i.e., differentiates performers on the basis of criterion measures), then the variable being measured should be taught. In the end, teaching and assessment of competence should include factors which cause effective use of both knowledge and behaviors in the classroom. For example, a teacher's ability to match teaching style with student learning style is an important teacher characteristic. It makes sense, then, to define, teach, and assess the causal variables or factors that lead teachers to make such adaptations. This approach is in contrast to merely articulating, teaching, and observing the morass of possible and necessary adaptations that will need to occur. Behavioral observations may provide evidence of whether the teacher performs specific adaptations under specific conditions, but these observed behaviors may not generalize to numerous new situations encountered in the field.

Furthermore, if the determination of teacher competence is to depend upon performances of students, we must expand the range of student learning outcomes to include cognitive, affective, and other antecedent causal variables. For example, let us assume that an agreed upon behavioral outcome for students is their ability to solve problems involving Boyle's law (borrowing from Broudy again). It may be more important to instill and assess general critical thinking and problem solving skills than merely to teach the law and demonstrate its use in limited sample problems. It is important that the student identify situations where Boyle's law is relevant as well as apply the law effectively in those situations. Moreover, the student's achievement in general problem solving or critical thinking is more important for adult life than is the knowledge and sample applications of Boyle's law out of the context of life situations calling for its use. Measures of these more generic abilities are beginning to emerge. They appear to be sensitive to important changes in students' learning and they measure learning outcomes which are important to fair and comprehensive determination of which teachers and which teachercharacteristics really make a difference.

In order to see the importance of these measures more clearly, I want to present the ICA model more graphically. I am taking the liberty to contrast the ICA model with the one presented by Medley, Soar and Soar (1975). This is not to denigrate their model, but to expand upon it. It is the clarity, parsimony, and conciseness with which they presented their model that allows

me to do this. A brief presentation of their model and quoted explanation follows:[1]

A Simple Paradigm

It will be confenient to distinguish four different levels in the teacher's professional development at which the teacher may be assessed, as shown in Figure 1.

LEVEL I	LEVEL II	LEVEL III	LEVEL IV
Training Experiences	Teacher Performance	Pupil Learning Experiences	Pupil Outcomes

Figure 1. Assessment Levels in Teacher Education

LEVEL I refers to assessments of the training experiences the teacher has had: What courses has he taken? What modules has he attempted? Which ones has he mastered? Which has he bypassed on the basis of having demonstrated mastery of its objectives beforehand?

LEVEL II refers to assessments of the teacher's behavior while he is attempting to fulfill the role of a teacher. What kinds of questions does he ask in interaction with pupils? How does he organize his class for instruction? How does he determine the objectives of instruction?

LEVEL III refers to assessment of the behaviors of pupils under the guidance of the teacher being assessed - assessments of the experiences they have which we all know must form the basis for any learning that takes place. What kinds of tasks do the pupils perform, in or out of class? How much time do they spend in active participation in class discussions? How often does each child receive reinforcement and for what?

LEVEL IV refers to assessments of the outcomes of instruction - of those changes in behavior that it is the purpose of education to bring about. How well does the pupil read? What are his attitudes toward independent learning in adult life? What kind of a citizen does he become?

Each of these four assessment levels is represented in the diagram, joined by arrows representing lines of influence or of cause-and-effect. Thus outcomes (LEVEL IV) are seen as influenced by, in part the result of, pupil behavior (LEVEL III), of learning experiences the pupil has while in school; these experiences in turn are seen as at least partly determined by what his teacher does (LEVEL II); and, finally, the way the teacher behaves - "teaches" - is affected by the experiences he has had during training (LEVEL I). All of these things are, of course, strongly influenced by other factors not shown in the figure, e.g., community, school, pupil and teacher characteristics.

---

[1]This model and brief explanation is taken from Medley, Soar, and Soar, Assessment and Research in Teacher Education: Focus on PBTE, PPTE Series: No. 17, Washington, D.C.: American Association of Colleges for Teacher Education, 1975.

The whole enterprise of teacher education is, of course, based on the assumption that, despite these extraneous factors, the influences that are assessed at each stage are potent enough to have an appreciable effect not only on the level immediately following but on all subsequent levels. The concept of teacher effectiveness, in particular, is based on the notion that pupil learning outcomes (LEVEL IV) are affected by teacher behavior (LEVEL II). And the justification for the very existence of teacher education is the presumption that what happens to a teacher in training (LEVEL I) can somehow increase his effectiveness, that is, affect pupil learning outcomes (LEVEL IV).

Implications of the Paradigm

The very existence of intermediate Levels II and III suggests that the effects of training on teacher effectiveness are subject to attenuation and are, therefore, difficult to establish. Any impact on pupil learning (Level IV) of teacher behavior (Level II) must be achieved through pupil behaviors (Level III). Two teachers who behave identically will achieve the same outcomes only if their pupils also behave in the same way. If we attempt to relate teacher behavfor to pupil learning without paying some heed to what the pupils are doing, we should not be surprised to find that the correlations tend to be low.

In the same way, the relationship between teacher education (Level I) and effective instruction (Level IV) depends on what teachers and pupils do in Levels II and III. We must train teachers to behave in such a fashion that their pupils will behave in such a fashion that the pupils will learn more!

The ICA model is presented in Figures 2, 3, and 4, where it has been juxtaposed or integrated with Medley, Soar and Soar's paradigm. The intention is to make more graphic the importance and possibilities of measuring critical causal intervening variables which are often stated as relevant to competence but rarely assessed or even operationally defined. This model reflects the notion that directly observable behaviors are insufficient indicators of successful outcomes at the pupil, teacher, and institutional levels.

Figure 2 demonstrates the notion that there are antecedent and often unobservable causal variables that affect teachers' learning experiences, teachers' performances in the field, students' learning experiences, and students' outcomes. Figure 3 demonstrates the phenomenon of intervening causal variables at the level of college-teacher interactions.[2] I have focused mainly on causal variables at the teacher and student level, but Figure 4 integrates all levels. It shows definitions, measures, and interactions of external behaviors and their antecedent and intervening causal

---

[2]At the program level there are causal intervening variables that are often discussed in terms of organization and management theory, but they are rarely assessed or accounted for in curriculum and program development. Rather, the focus remains on imparting knowledge and rote behaviors without a clear understanding of how even these outcomes can be achieved most effectively.

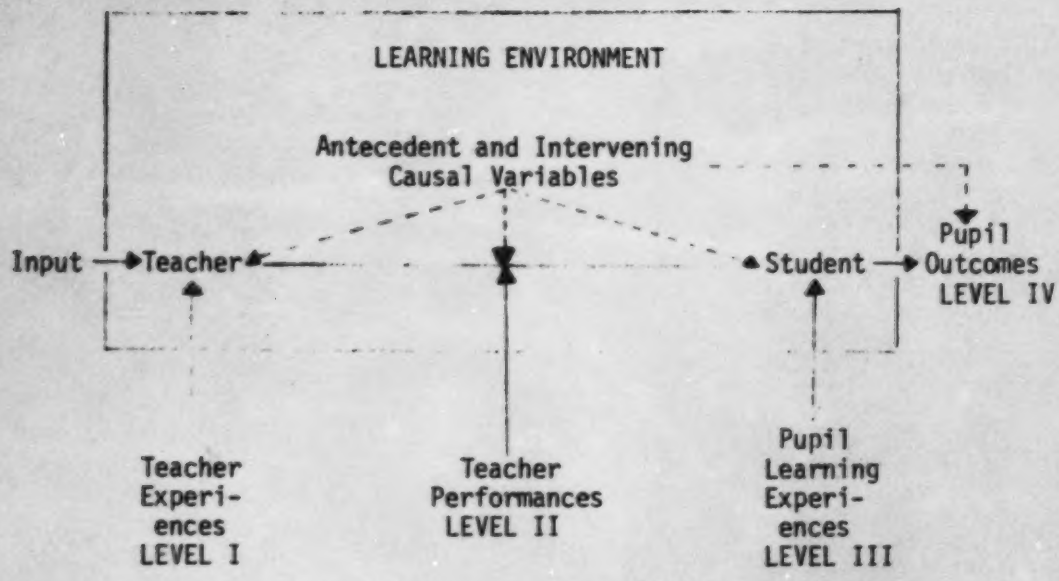


Figure 2. A Simple Model for Conceptualizing Causal Variables in the Teacher-Student Relationship

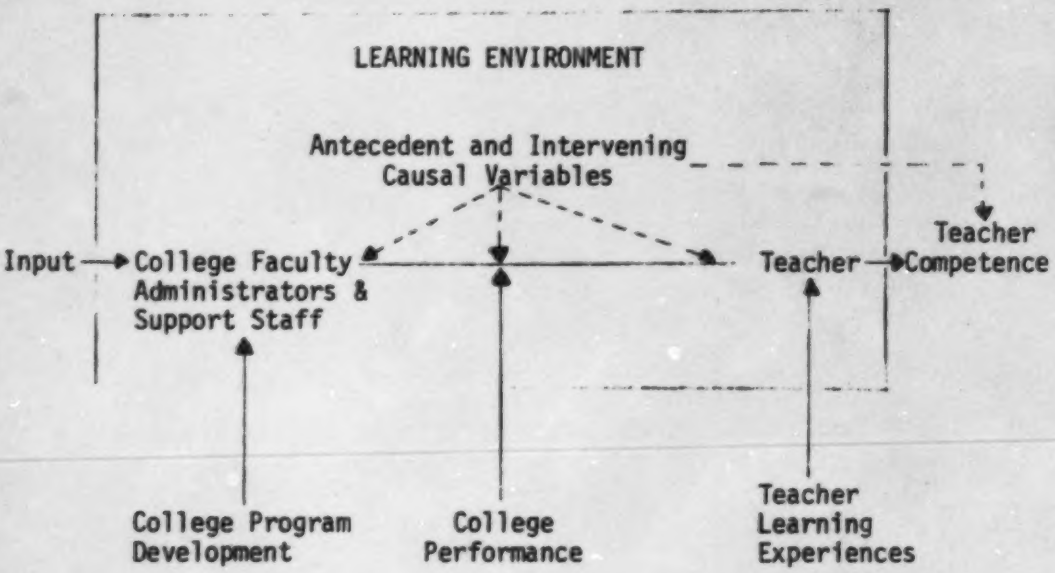


Figure 3. A Simple Model for Conceptualizing Causal Variables in the College-Teacher Relationship

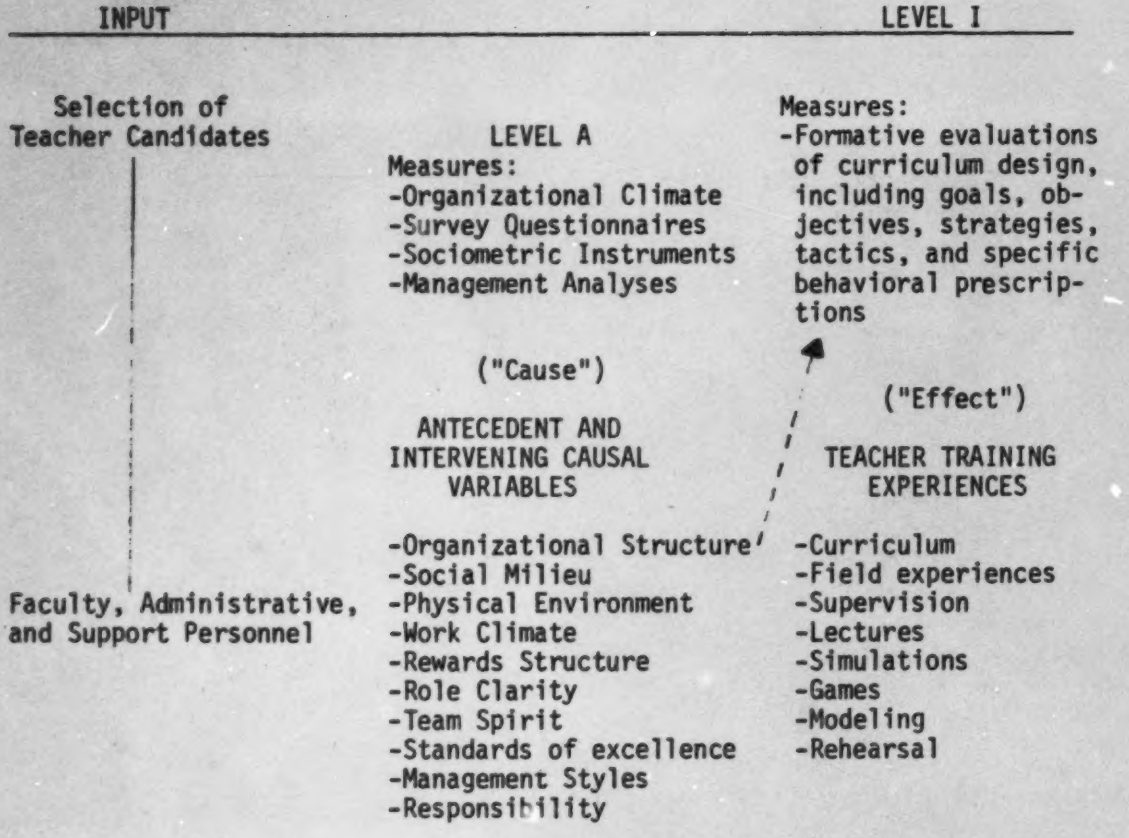


Figure 4. A Model for Conceptualizing Antecedent and Intervening Causal Variables of Institutional, Teacher, and Student Behaviors.

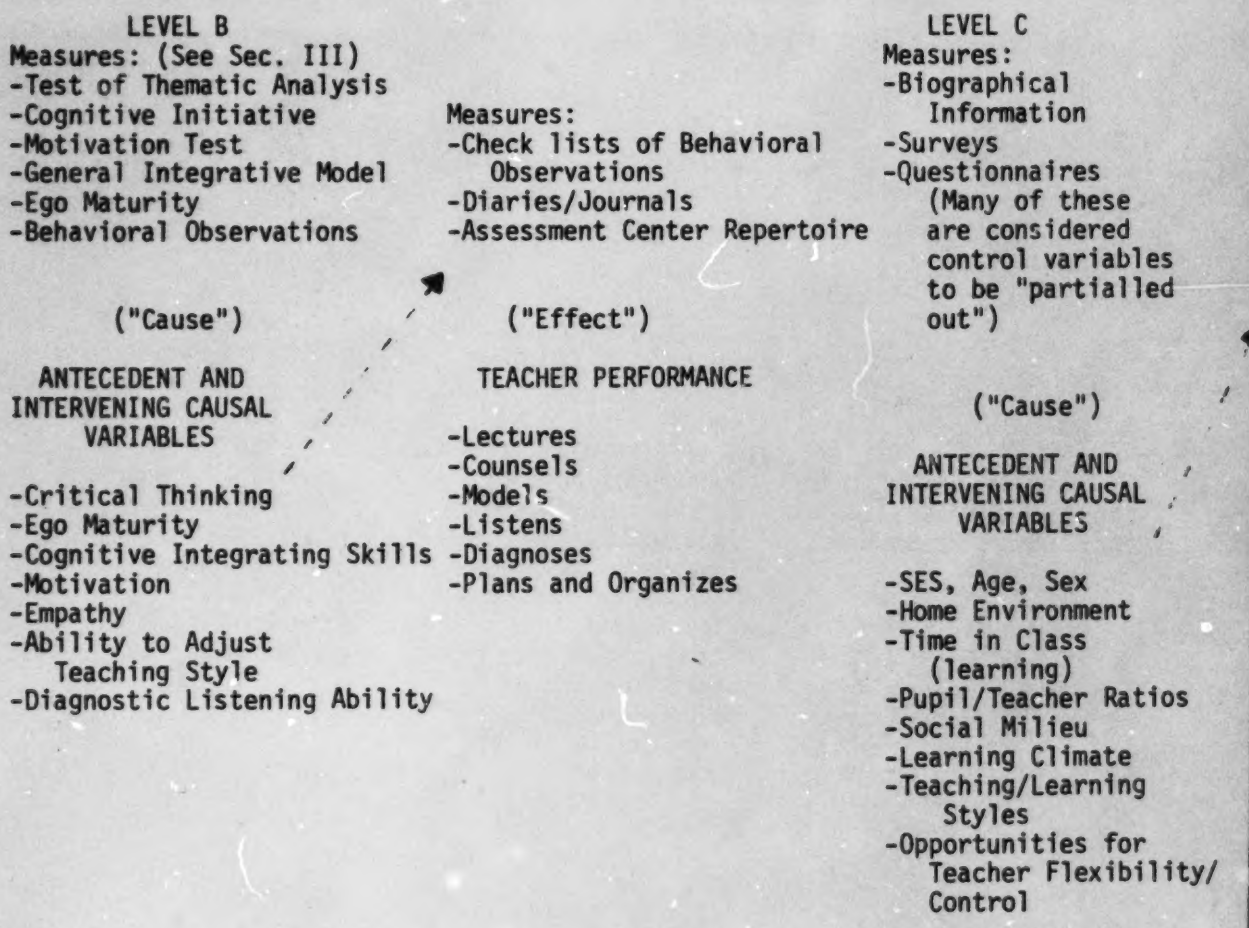


Figure 4. A Model for Conceptualizing Antecedent and Intervening Causal Variables of Institutional, Teacher, and Student Behaviors (page 2)

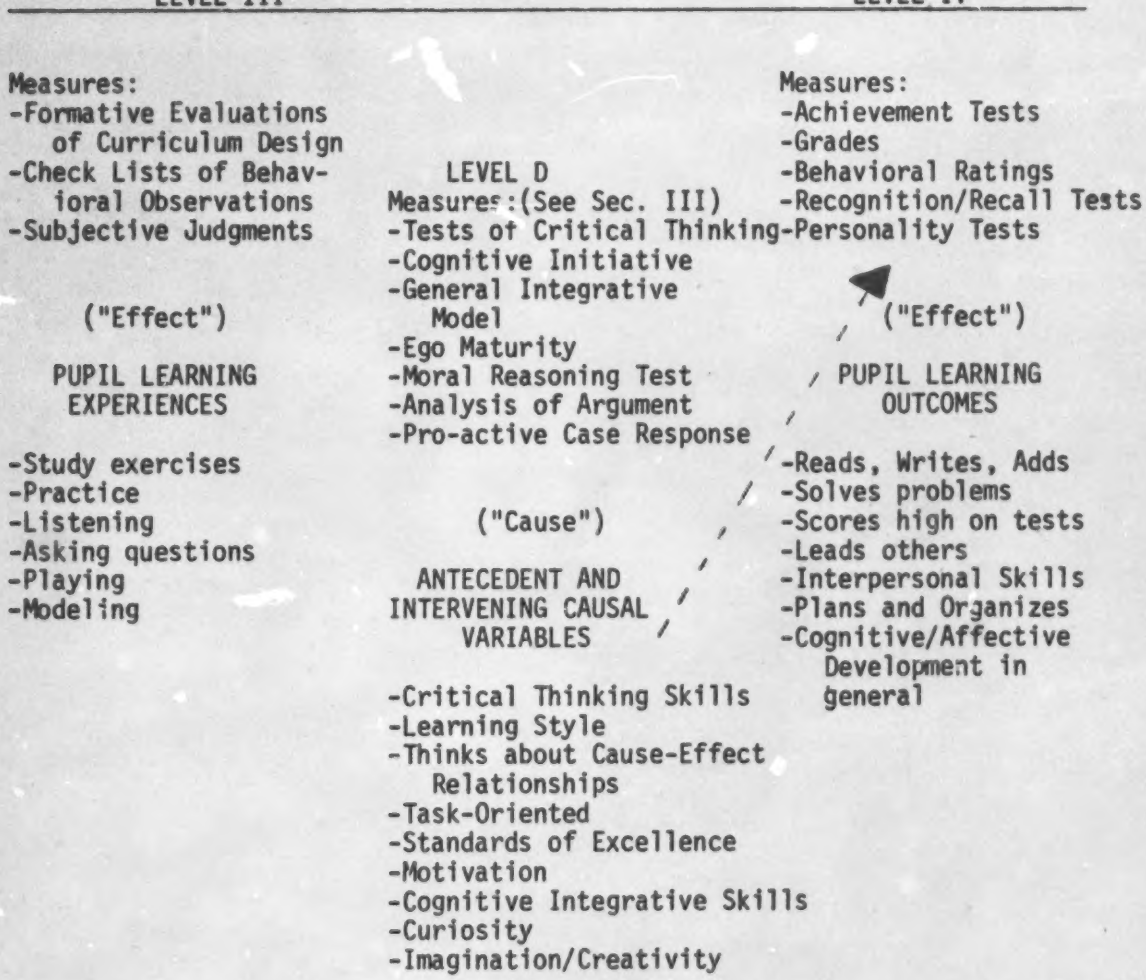


Figure 4. A Model for Conceptualizing Antecedent and Intervening Causal Variables of Institutional, Teacher, and Student Behaviors (page 3)

variables. These causal variables are presented in levels A through D. Their behavioral manifestations are presented in levels I through IV. Sample measures of both types of variables are suggested at each level.

It may be the more generic causal factors of competence that should (1) at least partially comprise competency definitions, (2) be taught in teacher education colleges, and (3) be measured in valid assessments of competence. This model would consider causal variables as legitimate learning outcomes as well as intervening variables.

Causally related criterion measures may be critical to teacher education because they provide the means for making better empirical links between teaching performances and student outcomes. This is so because these types of instruments appear to be more sensitive to short term and/or more relevant changes in student learning than standardized achievement tests.

The use of these types of measures can better clarify our understanding of what constitutes effective teaching (e.g., critical thinking, problem solving, the ability to implement effective solutions) because these measures can be made amenable to construct validation and empirical predictive validation. These validations are being demanded by researchers to understand better the meaning of measures, and they are becoming more strongly advocated by EEOC guidelines. They are also being required more frequently by our courts of law. It is because of these methodological and legal requirements for rigorous validation that behavioral checklists will ultimately have to give way to more relevant and objective measures.

Because causal related criterion measures can help clarify our understanding of what constitutes good teaching, they have important implications for curriculum, training and the modification of causal factors that underlie the behavioral performances which partially constitute effective teaching.

To assume that intervening or generic causal variables cannot be measured and are therefore unimportant is to take an unprofessional position with regard to competence definition and assessment. Researchers like McClelland, Winter, Stewart, Klemp, Kolb, Rosenthal, and others are attempting to develop valid measures of underlying causal factors.

A sample application of the ICA model comes from research with human service workers in the State of New York. McClelland and Dailey (1974) discovered that several factors were responsible for differentiating superior human service workers from average ones. These factors included (1) faith in the ability of people to change their behavior; (2) the ability to observe and diagnose human problems; (3) the ability to arrive at realistic, non-judgmental goals with clients; (4) imagination in thinking of solutions to problems necessary to balance clients' needs with department policies; (5) persistence in pursuing solutions; (6) the ability to remain task-oriented under stress; and (7) the ability to get people to work together in a variety of liaison activities.[3]

---

[3]Measures of these factors were developed and have been validated with other human service delivery personnel in the Massachusetts Civil Service Systems (McClelland and Klemp, 1975), and in Navy training centers (Klemp, personal communication).

In contrast, if one accepts the job description of a human service worker as a statement of behavioral objectives, the requirements for competence appear strikingly different. They describe only the behavioral manifestations of the seven factors listed above, such as "work in a counseling relationship with clients," "establish rapport," "make informed referrals when necessary," "make telephone and personal contact with agencies and services for and about clients," "have knowledge of specific community environment," "process benefits for clients within frameworks of department policies," and "make home visits." This is only a partial list. The point is there are few if any indicators of the abilities and other characteristics required to perform these tasks competently. Thus, it was more meaningful to discover and measure the generic causal variables that led to successful behaviors than it was to discover, endlessly document, and rate the specific behaviors themselves.

These findings are interesting because of the way they parallel those of education researchers. For example, Gage (1976) found that enthusiasm of teachers and teachers' acceptance of students' ideas yielded positive relationships with student achievement. Organization of teachers' presentations and teachers' use of "higher-order questions," however, were not related to student knowledge or comprehension. Yet, often, the latter two types of variables are included in competency statements. Furthermore, they are taught and assessed as if they were valid indicators of competence.

Clearly, our intuitions about what constitutes effective teaching must be empirically verified. And we must not eliminate the verification of variables which are intuitively important but difficult to measure. Such variables may constitute the very heart of competence. Measures of these variables that are objective, construct validated and empirically linked to successful teaching might provide the means for clarifying competency definitions and the means for validating these competencies to the satisfaction of everyone concerned.

Let me summarize what is needed to establish criteria for developing and selecting relevant cost-effective and valid assessment instruments before presenting some examples of new instruments.

(1) One need is for new conceptual frameworks to guide our thinking and processes in identifying, operationalizing and measuring competence. These conceptual models must emphasize the underlying causal variables that lead to observable behaviors rather than to new ways of simply cataloging and oversimplifying the morass of complex behaviors involved in the teaching process.

(2) Better techniques are needed for developing measures which tap these broad-based, generic, or generalizable causal factors. They must emphasize the quantification of outcome criteria so that educators can rigorously and meaningfully validate these measures. They must emphasize new methods of assessing performance rather than the predominantly subjective observational checklist techniques or the passive and respondent methods (such as standard paper and pencil knowledge recall tests) now in use.

(3) Practical methods for validating new measures are necessary so that institutions of higher education understand the meaning of their assessment measures and techniques. These methods must include construct validation.

(4) New measures must be referenced to criteria which reflect requirements for teaching success if the real meaning of measures is to be relevant. These relationships must not be mere correlations between observable behaviors during assessment exercises and successful behaviors on the job, but they must reflect causal links between learning and success.

(5) Measures are needed which (a) are sensitive to short-run changes, (b) provide useful feedback about the progress students are making toward their learning goals and (c) enable teachers and administrators to develop and evaluate better curriculum, teaching techniques, and supportive mechanisms for developing competent teachers.[4]

Prototype Measures of Liberal Arts Learning Outcomes and Other Competencies Related to Teaching

This section will describe some prototype measures which ICA has developed or utilized in liberal arts higher education programs and in other public and private sector organizations. Since they were developed or utilized in a variety of educational and occupational settings not all will have applicability to teaching competencies.

These descriptions were prepared by Dr. George Klemp, the Director of Research at the Institute for Competence Assessment as part of a proposal to develop a consortium of liberal arts competency-based colleges. The specific relevance of these measures for teacher competencies has not been articulated. Therefore, the present audience might find it meaningful to examine them in terms of understanding the types of measures ICA's empirical operational model develops or selects. Nonetheless, some of these measures are potentially quite relevant to teacher competencies, and these or similar measures should begin to be utilized in teacher education programs. ICA hopes to expand its repertoire of relevant measures for assessing competence over the next three years. This repertoire will be catalogued and disseminated. Aggregate data from their use will be analyzed and made available to the public.

By way of introduction, ICA discussed in its proposal the need for new measures which (1) are sensitive and relevant to important learning outcomes of liberal arts educators; (2) have general significance to a wide variety of career and life outcomes; (3) have practical utility for educators; (4) are methodologically and technically innovative, e.g., utilizing operant rather than respondent behaviors; and (5) are quantifiable and thus amenable to rigorous determination of reliability, validity, and meaning. Using these criteria ICA had developed innovative measures which attempt to answer the need for more "proactive" (operant) measurement techniques to assess the factors of process, integration and implementation.

---

[4]The issue of establishing criterion levels or performance standards has not been discussed in this presentation. A brief discussion of this problem adapted from Pottinger and Klemp (1976) is presented in Appendix C. This discussion further argues against simplistic assumptions of additivity and/or linearity in assessing specific behaviors to determine overall competence.

The purpose of this section is to present information about particular instruments which have been designed to measure competency-based outcomes. A subset of these measures is discussed in depth, and data relating these measures to academic and real world outcomes are presented. For the sake of clarity, and consistent with the competency-based orientation toward outcome-relatedness, the measures described below are organized according to three outcome domains: cognitive, effective and social outcomes.

Cognitive outcomes. Measures in this domain assess characteristics purportedly measured by traditional tests of mental ability, aptitude and knowledge. The differentiating characteristic between ICA's measures and traditional tests is that ICA's measures are based on the idea that the test-taker should provide all the information necessary for adequate and appropriate response to a problem on a test, as opposed to merely selecting from a set of prepared alternative responses.

Effective outcomes. Variables measured in this domain are directly translatable to behavior patterns required beyond the world of academia. This category is derived from White's (1959) term "effectance," which means positive, goal-directed and productive interaction with and influence on the environment.

Social outcomes. These measures assess areas of interpersonal competence which often facilitate the fruition of cognitive and effective dimensions of competence in life. They take into consideration the attitudes, values and orientations toward others which moderate life goals and the means for achieving them.

Measures of Cognitive Outcomes

1. Critical Thinking. The ability to analyze new information and to synthesize new concepts based on this information reflects the ability to integrate information into one's own cognitive structure. As the cognitive structure grows, so does the ability to think critically, to make a cogent argument and to reason inductively; thus, the test of Thematic Analysis is a measure of cognitive development. The test takes the form of two sets of stories which an individual is asked to compare thematically. This "thematic analysis" is scored according to twelve categories of critical thinking and a total score is derived. This scoring system is reliable, efficient and cost-effective. Each scoring category is a logical and independent dimension of critical thinking skill.

This test, developed by Winter (1973), is distinguished from other measures of critical thinking skills in that it demands the test-taker to actually produce critical arguments, rather than to simply recognize the critical elements of arguments presented to him. This instrument can be used to chart a student's progress in learning this skill. Alternative versions of the test have been developed to assess both the quality and structure of critical thinking.

Recent studies undertaken to assess the effects of the college experience upon undergraduates at Wesleyan and Harvard Universities (McClelland, 1976) show that seniors score higher than freshmen on this measure. It is important to note in this context that many so-called "cognitive" tests do not reflect the improvement in students' skill over the course of a four-

year college experience. When one examines firsthand the responses to the test of Thematic Analysis, however, it is not only clear that critical thinking skills improve with college, but that the scoring system for this test is intuitively satisfying in the ground it covers.

Under an ICA contract with The Fund for the Improvement of Post-secondary Education, Alverno College began to administer the test of Thematic Analysis to incoming freshmen along with other measures, including the Watson-Glaser test of critical thinking (a respondent measure), while the test of Thematic Analysis requires students to demonstrate critical thinking ability (an operant measure). An analysis of the data showed that the Watson-Glaser and Winter's measure of critical thinking were somewhat correlated, but only the test of Thematic Analysis was uncorrelated with respondent measures of other unrelated abilities. Those results speak favorably for Winter's measure as an uncontaminated test of critical thinking skill.

2. Learning Styles. A successful worker is distinguished not so much by a single set of knowledge or skills, but by the ability to adapt to and master the changing demands of one's job and career: that is, his ability to learn. Continuing success in a changing world requires an ability to explore new opportunities and learn from past successes and failures. Kolb's Learning Styles Inventory (1971) is a measure of individual learning styles which affect decision-making and problem-solving. The four styles, Concrete Experiential learning (CE), Reflective Observation learning (RO), Abstract Conceptualization learning (AC), and Active Experimentation learning (AE), when present in equal proportions, indicate the type of person who is able to involve himself fully, openly, and without bias in a new experience (CE), can reflect on and observe these experiences from many perspectives (RO), is able to create concepts that integrate his observations into logically sound "theories" (AC), and can use these theories to make decisions and solve problems (AE) (Kolb, 1973).

Extensive data have been collected on this measure in both college and postacademic settings (particularly the world of business). Kolb and Goldman (1973) have documented the utility of the Learning Styles Inventory for predicting major areas of undergraduate specialization and graduate school plans among M.I.T. undergraduates. The better the match between a student's learning style and the major subject area of the student's choice, the greater the tendency for students to place high importance in pursuing a career in that area, to perceive their workload as light, and to involve themselves with important peer groups, and the lesser the tendency for students to experience disaffection with their social and academic experience.

More recent work involving the analysis of administrative and technical support positions in the Division of Civil Service, Commonwealth of Massachusetts, identified "the ability to learn from experience" as a key to worker success. The Concrete Experience (CE) scale of the Learning Styles Inventory was found, in fact, to be significantly correlated with superior performance in this category of work, involving over 15 job titles (Klemp, 1976).

3. Programmed Cases. Based on incidents called from in-depth interviews with criterion groups, programmed cases can be developed to test

for social learning and judgment. Versions of this technique, developed for the U.S. Information Agency and the U.S. Navy, consist of a series of incidents to which several alternative responses are attached. All of the incidents pertain to a particular individual, or "case." "Distractors," or the incorrect responses, are developed with the aid of expert judges. The cases are programmed in such a way that a person with good judgment, i.e., who does not make snap, impulsive judgments, will become more accurate in his choices of the correct alternative as he proceeds through the case.

The programmed case technology has two primary uses:

- diagnostic assessment of how one uses information in making decisions about others or predicting their behaviors, and

- examination of the process by which decisions/predictions are made, including the analysis of values, biases and preconceptions that interfere with veridical impressions of others and their situations.

These programmed cases are currently being used in psychological studies at Harvard as a measure of interpersonal learning. McBer's research interest in this technology has led to applications of programmed cases in the study of prejudice.

Klemp (1975) found that people who were exposed to cases about people whose race was unlike that of the reader were less able to predict the behavior of the person in the case than readers who were exposed to same race cases. Similar studies are planned to address the prejudicial effects of socioeconomic status and sex differences on interpersonal learning.

The direct application of programmed cases, other than personnel selection, has been in assessing the skills of human relations experts in the U.S. Navy. In a pilot study (unpublished) involving selected human resource training personnel whose performance level was known, a highly significant relationship obtained between the ability to accurately predict behavior in others, as measured by the programmed cases, and performance as a trainer in human resource management.

Other measures of cognitive outcomes, in prototype form, are the following:

4. Analysis of Argument. A test of the ability to argue for and against a controversial issue, and scored for the logical presentation of argument (Stewart, 1974).

5. Concept Formation. A test of the ability to identify and organize similarities and differences among objects into concepts.

6. Speed of Learning. A test of how quickly one can learn new material selectively - that is, to remember functionally important information.

7. Savings Score. A test of the ability to learn new material in a particular content area - to "save" new information in an area in which the student is already well versed.

8. Proactive Case Response. A test of diagnosis, judgment, and problem solving that involves response to a detailed situation, or "case."

Measures of Effective Outcomes

1. Diagnostic Listening. The Diagnostic Listening Test consists of a taped presentation, with slides, of interviews with various individuals typical of the people one might encounter in social service work. People who take this test listen to an interview or a brief statement by a particular individual on the tape, and are then asked some questions about what has happened, what the person is like, and what they would recommend for the person. This test requires listening, observing and judging skills which have been found necessary in human service work.

There are two subscales in this test. The Casework Subscale, consisting of 42 items, is made up of four interviews and after each of them the person taking the test is asked to answer questions and to make judgments of a multiple-choice answer sheet. The Positive Bias Subscale, consisting of 39 items, shows to test-takers three slides of clients of different sex and race with accompanying brief monologue. After each of these presentations, the test-takers are required to rate several adjectives as "does describe" or "does not describe" the client. An overall Positive Bias score is obtained by summing the number of positive yet realistic adjectives chosen. The Diagnostic Listening test measures faith in the client's ability to change, ability to observe and diagnose human problems, ability to set realistic goals, and ability to propose imaginative solutions.

Studies of human service workers in the State of Massachusetts have verified the usefulness of the skills tapped by the Diagnostic Listening Test in identifying better workers. The format of the test instrument is similar to interview situations in which workers are involved on a day-to-day basis. Both of the two subscales correlate with effective on-the-job performance as rated by supervisory consensus (McClelland and Klemp, 1974).

Introduction of Measures 2 and 3: Much research has been accumulated by McClelland (1958, 1961), McClelland and Winter (1971), Atkinson (1958), and others that shows that thought patterns are related to important kinds of behaviors. The Exercise of Imagination is McBer's version of the Thematic Apperception Test (TAT) which is used to elicit thought patterns of the test-taker.

An individual taking the test is asked to write narratives to pictures. Each of these narratives addresses the following questions about the pictures: What is happening? Who are the people? What has happened in the past that has led to the situation? What is being thought? What is wanted by whom? What will happen? and What will be done? The stories are then scored, according to a prescribed set of codes or rules, to uncover certain patterns of thought that are expressed in the stories. These scoring codes can be applied to any written narrative which addresses the types of questions mentioned above.

The link between thoughts and behavior has been repeatedly demonstrated

to be strong, as opposed to the link between attitudes and behavior. The attitude-behavior link is influenced primarily by situational factors. An attitude may represent a specific goal or objective, but such goals and objectives may change according to situational demands and constraints. However, whether a specific goal changes or not, the characteristic style with which any goal is attained is determined to a large extent by thought patterns which are relatively consistent within individuals.

2. Achievement Motivation. McClelland has shown in extensive research (1961) that people high in the need for achievement are practical and interested in efficiency - in short, they are good practical decision-makers. They are independent, good at evaluating information for its practical utility, and original in the sense that they keep looking for better ways of doing things. For instance, they make good career decisions and regularly achieve greater success earlier in their careers. In a recent Harvard University longitudinal follow-up study, freshmen n Ach (need for achievement) scores correlated with "early success" in various fields 14 years later (McClelland, 1976).

In the world of business, studies have shown that achievement motivation is highly related to small business success, success in sales, and performance in the role of entrepreneur (McClelland and Burnham, 1976). The need for achievement, the desire to do things better than anyone else, is particularly great among scientists and others who work against a self-imposed standard of excellence. People low in achievement motivation generally do not exhibit planning or goal-setting behavior, nor do they weigh the risks they take against expected gain. The habits of behavior in such persons may not be advantageous to success in school or in many kinds of careers. But McClelland (1965) has pointed out that people can be taught to behave in ways that are reflected by the achievement motive, and so the gap between successful performance in certain academic and work settings may be effectively bridged.

3. Self-Definition/Cognitive Initiative. Self-definition/cognitive initiative is a general characteristic of an individual which encompasses the way one thinks about the world and himself, the way one reacts to new information, and the way one behaves. People with this competency are not only able to think clearly, but also to reason from the problem at hand to a solution, and to propose and take effective action on their own. Such competence is characteristic of people who think in a rational, systematic way on their own, and who can anticipate problems before they arise. In short, it might be said that people who are high in this characteristic are able on their own to see things clearly, to understand the causes of events, to reason from problem to solution, and to take effective action to solve problems. For example, the self-definition score has been quite useful in distinguishing between women who pursue careers following college and those who do not (Stewart and Winter, 1974).

A longitudinal study involving freshmen women at Alverno College begun by McBer with FIPSE funds, will track Self-Definition/Cognitive Initiative during the four-year college experience. The preliminary data on this measure show that it is uncorrelated with other measures of college-entry knowledge, skills, and abilities. It is therefore considered to measure a unique dimension that, because of its known predictive validity regarding the success of women in careers, is a particularly important measure in a competency-based assessment system.

Other measures of effective outcomes, in prototype form, are the following:

4. Socialized Power. A measure of whether a person is motivated to express or increase his own power for the good of the self or for the good of others.

5. Stage IV Power. A recently identified measure (McClelland, 1975) of a concern for doing one's duty, that is, to be an instrument of a power which extends beyond the self.

Measures of Social Outcomes

1. Nonverbal Sensitivity. This test, developed by Rosenthal and his associates at Harvard University (1974), consists of 40 brief voice segments on tape, all of which have been altered to obscure the words. There are two subscales to the test: The RS Subscale, made up of voice segments that are randomly spliced and reassembled, and the CF Subscale, made up of segments which have been electronically filtered so that the words are unintelligible, but the intonation patterns remain. A sample item would consist of a speech segment followed by a question; e.g., "Does the segment represent someone helping a customer or criticizing someone else for being late?" Rosenthal has documented some promising criterion validity for the PONS Test. High scorers on this test exhibit the following characteristics:

- they represent warmer, more honest and more satisfying peer relationships;
- they have been rated by peers and/or by teachers who know them well as being generally more sensitive in interpersonal situations; and
- they were found to be functioning more effectively in the social and intellectual areas of the California Personality Inventory.

This test, which requires less than 10 minutes to administer, has been found to predict successful performance in administrative and human service jobs, which require that the worker have "empathy," or the "ability to read between the lines" in the performance of the job (Klemp, 1976). Navy personnel involved in race relations work also have been found to score higher than the general population on this test, and the personnel who are more successful on the job also score higher than their less successful counterparts.

2. Moral Reasoning. This test is based on the research in moral development by Lawrence Kohlberg at Harvard (1970). The test consists of a series of paragraphs which describe complex situations in which the actors are forced to choose among several moral courses of action. The task of the applicant is to write a paragraph to justify the alternative that the applicant feels is the best one on moral grounds. The essay answers are scored according to a thematic analysis developed by Kohlberg, and are interpreted according to a schema containing six levels of moral development:

Stage 1: Orientation to obedience and punishment-deference to a superior power or to trouble-avoidance.

Stage 2: Orientation to action that is satisfying to the needs of the self.

Stage 3: Orientation toward approval and to pleasing and helping others.

Stage 4: Authority and social order maintenance orientation-"doing duty" and showing respect for authority.

Stage 5: Orientation to duty defined in terms of a contract, general avoidance of violation of the rights of others, and majority will and welfare.

Stage 6: Orientation to high principle or conscience.

The conceptual categories on which the test is based have a high degree of validity as constructs.

Some recent work in the medical profession has related Kohlberg's work to the practice of physicians. High relationships exist between a physician's level of moral development and whether he will withhold or pursue treatment, the degree to which he considers the patient in the context of his family, and overall ratings of physician performance. These results show the Moral Reasoning Test to be predictive of important kinds of behavior in work which requires a good deal of value judgment. As the study of one's own values is becoming a part of what many competency-based programs wish to offer their students, Kohlberg's stage orientation to moral development is offered as an important component to this educational experience.

Other measures of social outcomes, in prototype form, are the following:

3. Affiliation Motivation. A measure that reflects one's concern for the compassionate quality of relationships, helpful in building good working associations with friends and colleagues.

4. Social-Emotional Maturity. A scoring system of written narrative built upon a stage theory of maturity (Stewart, 1973) that makes sense both theoretically and intuitively. Table 1 presents a summary of this section.

A General Integrative Model

Of the tests and measures outlined in the preceding section, none is expecially useful as a diagnostic or assessment tool outside of a systematic approach to understanding the integration of the many skills that are required for success in life and work. The measures may be important pieces to the puzzle, but one cannot tell from pieces alone what the whole individual will look like. From the standpoint of competency-based education, it is the meaningful integration of life skills that is important as an outcome of the educational experience. The General Integrative Model is one way of expressing this value by involving several different measures in a system that can be used to assess student competence.

Table 1

Competency Based Measures and Their Developmental Status

	Prototype	Validated	Data Available in Academic Setting	Data Available in Real World Setting
Cognitive				
1. Critical Thinking		X	X	
2. Learning Styles		X	X	X
3. Programmed Cases		X		X
4. Analysis of Argument		X	X	
5. Concept Formation		X		
6. Speed of Learning	X		X	
7. Savings Score	X			
8. Proactive Case Response	X			
Effective				
1. Diagnostic Listening		X	X	X
2. Achievement Motivation		X	X	X
3. Self-Definition/Cognitive Initiative		X	X	X
4. Socialized Power		X		X
5. Stage IV Power	X			
Social				
1. Nonverbal Sensitivity		X	X	X
2. Moral Reasoning		X	X	X
3. Affiliation Motivation		X		X
4. Social-Emotional Maturity		X	X	

Such general competencies as the ability to cope with new problems, to find appropriate solutions, and to take the correct action steps can be considered in such a model. Figure 5 outlines one approximation to a systems approach that involves an integrated set of measures in a particular problem area, allows assessment at various junctures in the system for diagnostic purposes, and that also serves as a model for learning new skills through feedback in one's own performance. This particular version of the General Integrative Model requires an individual to demonstrate the following abilities:

- to observe;
- to extract relevant information;
- to analyze and integrate this information;
- to ask appropriate questions;
- to process new information in response to such questions;
- to utilize this information and one's knowledge in making sound and logical recommendations;
- to develop main and contingency plans;
- to set meaningful goals; and
- to feed back this new information into the process for better problem analysis and solutions.

This model is not a measure per se, but a collection of measures logically ordered, to assess problem solving skills. The progress from stage to stage in the model presents the students with subproblems to solve, e.g., what new information to seek, what conclusions to draw, and what decisions to make derived from the information gathered at a given time.

This particular model emphasizes cognitive skills, but other models can be developed that deal in different areas of competence. For example, the U.S. Navy, in their Human Goals Program, is striving to implement a training model that uses as input tests of achievement, affiliation, and power, programmed cases, learning styles, and sensitivity to nonverbal communication. By using this model, the Navy seeks not only to assess and diagnose, but to develop curriculum aimed at more effective preparation of their personnel for work.

Characteristics and Advantages of Competency-Based Measures

This section pertains particularly to the measures outlined above, but may also be considered to be the hallmark attributes of competency-based measurement in general.

1. These tests require the person being tested to be proactive, not just reactive (i.e., one has to generate responses which can be scored for

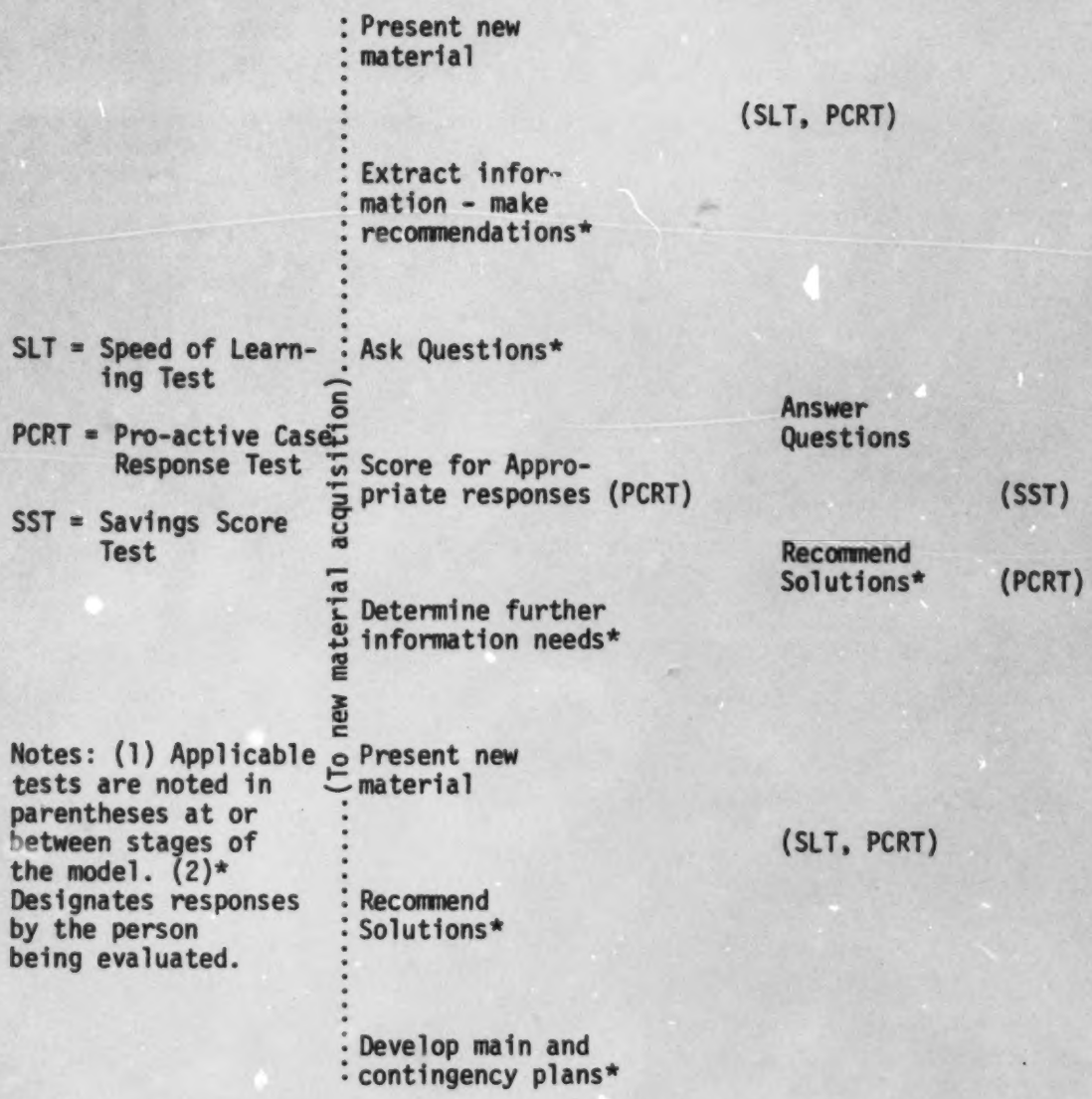

Present new material
(SLT, PCRT)
Extract information - make recommendations*
Ask Questions*
Answer Questions
Score for Appropriate responses (PCRT)
(SST)
Recommend Solutions*
(PCRT)
Determine further information needs*
(To new material acquisition)
Present new material
(SLT, PCRT)
Recommend Solutions*
Develop main and contingency plans*
SLT = Speed of Learning Test
PCRT = Pro-active Case Response Test
SST = Savings Score Test
Notes: (1) Applicable tests are noted in parentheses at or between stages of the model. (2)* Designates responses by the person being evaluated.

Figure 5. A General Integrative Model
(One approximation)

their appropriateness to real life situations). Thus, the test-taker goes beyond recognizing answers out of context. In the general model, if timing of questions or recommendations is a critical aspect of problem-solving, then this time variable can be programmed into the model as well.

2. The tests are efficient since they can be given to groups as well as to individuals. Their efficiency and economy should substantially reduce the operational costs of current assessment procedures which require vast amounts of time, people and other resources.

3. These instruments foster equity in the assessment process, since they can be objectively and reliably scored according to the empirically validated coding systems. This is an important advantage since current methods of using juries, panels, or other groups to evaluate are not only inefficient and uneconomical, but are also vulnerable to all the vagaries of subjectivism.

4. The scores can be standardized with reference to criterion groups of which a student is preparing to become a part.

5. Many of these tests tap the competency of "learning how to learn" in a content area. This is one of the most important competencies people can develop because throughout their lives they will be faced with the problem of learning new things in selected areas.

6. These tests are much less threatening and anxiety-producing than traditional tests of recall or recognition, which because of their properties only contribute to the fear of failure so prominent in nontraditional students.

7. A number of variations of these tests and the General Model can be developed to add flexibility for administrators, e.g., they lend themselves to video taping, written or oral answers, individual or group testing, etc.

8. The majority of these tests have face validity. Educators and students recognize that the skills and abilities being demonstrated are applicable to general life skills.

9. Empirical and construct validation with various occupational life skills outside of academia means that the competencies required for successful performance beyond the academic program can be established as the target of the learning process.

10. The models and tests can be validated with a variety of nonoccupation-specific populations. Some tests and models developed are noncontent-specific such that a competent person with little formal education can demonstrate competence as an analytic thinker, information processor, and a proactive initiator of appropriate solutions. The test format is easily followed and is attractive to those who are test-anxious in traditional test settings.

11. These measures can serve as pedagogical devices as well as assessment instruments, since practice in dealing with the information and com-

ponent competencies necessary to solve the test problems is a direct way of learning. The instructor and student alike can easily locate and analyze weaknesses and strengths of an individual in exercising component skills. Thus, these measures can serve as diagnostic and guidance tools for supplementary curricular modules.

12. One need not take a particular course or go to a particular college in order to attain competence in the generic skills and abilities measured by these assessment tools.

APPENDIX A

BRIEF GUIDELINES FOR TEST DEVELOPMENT AND UTILIZATION

(From Pottinger and Klemp, 1976)

These guidelines are partially based upon three aspects of assessment which have been addressed elsewhere by Pottinger (1975):

- The identification and definition of competencies relevant to life and work outside of academia;
- Instrumentation, techniques, and processes of evaluation that provide reliable and valid measures of these competencies; and
- Standardization, and/or establishment of levels of performance necessary and sufficient for awarding credentials.

A. "New" competencies must be identified and operationally defined.

There are many outcomes of the learning experience that have greater validity than grades in school as a basis for awarding credentials. Those that have been identified have been accepted as being important and meaningful in establishing a person's competence, yet many academicians have not sought to operationalize, measure, and award credit for many of these learning outcomes. There are many other criteria than traditionally-rewarded scholastic achievement that are important as competencies in the practical work, and most of them are as yet unidentified.

B. New competencies should have general significance to a wide variety of career and life outcomes.

Competencies cannot be meaningfully defined by a seemingly endless reduction of specific skills, tasks and actions which ultimately fall short of real-world requirement for effective performance. In fact, the more essential characteristics for success often turn out to be broad or generalized abilities or characteristics which are sometimes more easily operationally defined and measured than an array of specific "subskills" which do not add up to general competence.

C. New definition of competencies and measures developed for their assessment should be easy for faculty and students to understand and use.

New competency definitions should be readily recognizable as important, and related assessment techniques and instruments should be easy for faculty and students to understand. It is necessary to guard against competency definitions and measures that are so complex, or trivial, or esoteric that students and faculty can neither understand them, nor accept them as meaningful and useful. In other words, educational goals should not be rendered unintelligible, and assessment procedures and instruments should not mystify the process of evaluating students' progress.

D. Competencies should be empirically linked to external realities.

Many educators assume that such things as the ability to master new bodies of knowledge quickly and effectively, to analyze and solve problems, to develop new skills efficiently, and to utilize knowledge are prerequisites for individuals if they are to take advantage of life's opportunities

and surmount its difficulties. What is missing are the measures of these general abilities, which are related to important life outcomes. Only when we know what makes the difference between adequate and inadequate performance, based on empirical analyses of professions and other life activities, will we be able to develop or improve such measures, clarify new competencies, and establish credentials of demonstrated value.

E. The discovery of new ways of measuring abilities (competencies) is needed.

The measurement technology must be innovative and new, not just a new name for traditional procedures. Paper-and-pencil (objective) tests, due to method variance, correlate better with each other than they do with performance criteria. If postsecondary education is to break out of this closed circuit, different approaches to testing must be sought in areas such as learning, critical thinking, problem solving and other newly defined competencies. Measures of competence must require that the test-taker generate appropriate learning outcome responses. The primary learning objective of education is not to help an individual select from among a set of predetermined alternatives. Rather, it is to enable a person to know how to reason; how to marshal evidence for or against an hypothesis; how to see similarities and differences in objects, ideas, and events; how to partial out crucial information from the trivial; and how to integrate these skills with purpose and meaning. Multiple-choice tests do not and cannot measure these abilities. And the behavioral observation/documentation appraches that are popular in experiential learning assessment do not allow these abilities to be measured with adequate reliability or validity.

F. Standards of performance for awarding credentials should acknowledge levels of performance required for entry into roles outside of academic settings.

The establishment of criteria or standards of competence is one of the most difficult problems to be addressed. In every case, where standards of competence are determined for new or for more traditional outcomes, appropriate levels should be established by sufficient empirical evidence to ensure that they will not be viewed as arbitrary. Many educators are satisfied with a priori judgments of what skills and levels of performance are adequate. It is startling to realize how much we accept the face validity of credentials and how little we really know about (1) the correspondence between the abilities and levels of performance that these credentials represent, and (2) what is needed for adequate performance in life's tasks. We need to develop better benchmarks for evaluating the standards and the offerings of postsecondary institutions. (For further comments about the establishment of standards and levels of performance, see Appendix C.)

G. New attempts to define and assess learning outcomes should not be guided solely by attempts to make them functionally-equivalent substitutes for traditionally assessed school achievement.

Competency-based education requires a different type of evaluation from traditional programs to the extent that learning outcomes differ in significant ways. For example, learning outcomes in CBE are often defined in terms of what a person can do, not merely in terms of what one knows. Furthermore,

> ...whereas in traditional programs evaluation is primarily linked to the credentialing process, in competency programs it is also used as a formative teaching tool. In other words, students are made aware of the criteria and standards for certification in a competency, and their progress is frequently measured so that help can be provided as necessary. Assessment that simply places students in a percentile or just discriminates between passing and failing is not adequate for competency-based programs. Formative diagnostic advice is needed - information that tells if the student is "real world" competent (Hodgkinson, 1975).

The temptation to restrict the development of new measuremement instruments, techniques, and procedures, in order to achieve comparability with those that have gone before, has great political appeal for making such innovations palatable to traditionalists. However, if institutional and credential reforms are to succeed, we need to move beyond the recognized limitations of traditional assessment systems.

APPENDIX B

DETERMINING THE MEANING OF MEASURES

As background to this discussion, we have already stressed the importance of, for example, changing the focus of assessment from merely asking for recall and recognition of content to measuring how one processes and utilizes this information. If assessment techniques are to have sufficient meaning and credibility for determining if students are adequately prepared for teaching, we can no longer be satisfied with content-valid tests. Construct validation must be determined. Furthermore, PBTE proponents have stressed the importance of creating criterion-referenced measures which are predictive or reflective of real world requirements for successful teaching. The following discussion (Pottinger and Klemp, 1976) is a further elaboration on the necessity for construct validation and empirical linkage of measures to obtain maximum meaning of what is being assessed.

Messick (1975) has argued that, until measures have been construct validated, they lack the meaning essential to utilizing them as instruments of general education theory. McClelland (1973) further argues that, until construct validated measures use relevant real world events among their criterion referents, their value in assessing preparedness for work and life is limited. Educators have often failed to pay attention to construct validity because they "view desired behaviors as ends in themselves with little concern for the processes that produce them or for the causes of the undesired behaviors to be rectified" (Messick, p.959). In other words, "construct validity is not usually sought for educational tests, because they are typically already considered to be valid on other grounds, namely, on the grounds of content validity" (Messick, p. 959).

In short, educators have traditionally been satisfied with knowing that the content of tests adequately sample a class of situations or subject matter. Messick (1975) argues that content validity does not provide an evidential basis for interpreting the meaning of test scores, and McClelland (1973) argues further that the interpreted meaning of scores that come from construct validation must be strengthened by tying these constructs directly to the world of events outside of academia.

The theoretical distinction between general education and competency-based education is that the latter requires an empirical and causal link between measurement responses and their meaning, as related to real world life outcomes. Most competency-based programs, however, merely correlate test responses with specific criterion-referenced outcomes (and many do not even do this) without discovering the underlying causes of these responses. Many educators make the mistake of thinking that if a test correlates with a behavioral criterion variable in the world of work or elsewhere outside of the academic world, one can develop competence by "teaching to the test." But this notion confuses correlation with causation, i.e., the fact that tests correlate with

observable criteria may only indicate the existence of a causal intervening variable which is really responsible for behavior and which has not been measured.[5]

Clearly the mandate for competency-based postsecondary education is to identify skills and abilities that produce (cause) desired outcomes; to develop curricula aimed at the acquisition of these skills and abilities; and to design and validate measures that are sensitive to the acquisition processes and are representative of the criterion outcomes. One should not consider curriculum development apart from assessment issues and neither should it be considered in the absence of identified valid performance criteria. Only when these conditions are satisfied does it make sense to "teach to the test."

The skills tapped by genuine competency-based tests (i.e., causally-related criterion measures) are largely independent of the content areas in which they are used. For example, the tests for thematic analysis, analysis of argument, problem solving, speed of learning, and other such measures described in the text test for generic abilities (competencies) which can be demonstrated in the context of any specific content area. These tests can be adapted to the natural sciences, social sciences, and humanities with equal facility; the content area does not determine the effectiveness of the test. We will always need tests of knowledge, but we also need tests of the way this knowledge is used. The measures discussed in the third section satisfy both of these criteria, which represent the essence of competency-based assessment.

A criticism leveled at the PBTE movement is that it is limited to preparation of "didactical technicians" (Broudy, 1972). A narrow correlational model of competence has fostered this notion, and this concern is legitimate to the extent that criterion validities depend exclusively upon specific task-oriented behavioral criterion references. Such validities for teacher education "are of sporadic interpretive utility" at best since they ignore the linking of test behavior to a more general attribute, process, or trait which provides an evidential basis for interpreting the processes underlying test scores. (Messick, 1975)

ICA strongly endorses this position but adds that construct validation is itself all too often limited in the types of referents it uses to provide meaning to test scores. Thus, we advocate a validation more heavily in the context of other constructs alone or

---

[5]For example, vocabulary is correlated with college grades. However, one would not go about improving college grades merely by increasing vocabulary. Doing well in school requires abilities for problem solving, utilizing new information, and other skills not measured by vocabulary tests. Vocabulary is merely a tool, and how it is used depends upon other abilities and characteristics of the individual. One cannot do well in school without a reasonably adequate vocabulary, but having a strong vocabulary will not guarantee success in school without its effective use.

"laboratory" behaviors. While Messick (1975) de-emphasizes criterion-referencing, he only does so (1) in terms of using criterion-referents outside of the context of construct validation and (2) perhaps in terms of criterion used as referents. Indeed, all validation is criterion-referenced. The difference in criteria (e.g., "real world" performance, other tests, or observable "laboratory" behavior) determines the extent to which the meaning of the test responses are general or specific and of theoretical or real world significance. A difference between McClelland's (1973) and Messick's point of view is McClelland's emphasis on choosing real world behaviors as opposed to tests (which typically tap respondent rather than operant behaviors) and laboratory behaviors, as criterion referents. Thus, criterion-referents constituted by a nomological network of life outcomes are consistent with Messick's argument. Espousing such referents differs from Messick's point of view only in terms of emphasizing their selection as criteria for construct validation, not in the validation procedures or concepts themselves. In other words, Messick's notion of construct validation theoretically would include criterion behaviors, but empirically there are differences in emphasis on the types of behaviors to be included. It is for the sake of this difference in emphasis, not theoretical differences, that we have isolated real world events or life outcomes as critical factors in determining the real meaning of tests.

The strength and future of PBTE rests on its ability to support the rigorous type of research analysis which involves construct validation based heavily upon real world teaching outcomes. Until we have identified the critical intervening variables in the causal chain between the educational experience and performance in the field, we will be legitimately faulted by critics who view PBTE assessment (and education) as too narrow in scope.

APPENDIX C

THE PROBLEM OF ESTABLISHING CRITERION LEVELS

OR PERFORMANCE STANDARDS

As the meaning of measures becomes established by construct-validation and empirical (criterion-referenced) links between education and the requirements of postacademic life, the question of what criterion levels of performance are necessary for granting credentials is made easier, perhaps, because concrete information exists with which educators can make sound judgments. Yet, the problem of establishing standards for levels of performance is a complex one because (1) this determination of appropriate levels of performance is dependent upon educators' goals for credentialing students, and (2) technical issues related to understanding the meaning of maximum levels of performance and the meaning of complex interaction of abilities probably necessitate highly subjective determinations of criterion standards.

With regard to the first point about determining standards of performance, Hodgkinson (1975) stresses the importance of asking good questions about the use and purposes of assessment. Sound judgment and planning are necessary to avoid proceeding with evaluative decisions based on ambiguous criteria, standards and/or levels of outcomes. These questions must include: Who establishes criteria or standards - an external auditing agency, a faculty member, the institution? What is the reference group with which one will be compared - performers in the real world, students in past years, other students currently being evaluated, one's own past performance, an ideal student? What is the proper method of comparison? - norm-referenced tests, criterion-referenced tests, behavioral measures, narratives (e.g., portfolios, diaries of past experience), unobtrusive measures, etc.? What is the nature of the standard - job performance in the "real world," individual growth and development, ideological ideals of performance, standardized scores? What is the function of the standard - to select or reject people, to improve performances, to admit students to professional schools or jobs?

If these questions are asked and the answers are concrete, specific and meaningful, a student should know who is judging him, how he will be judged, the nature of these judgments, the objectives related to them, and how well he must perform to meet those objectives.

With regard to the second point about determining standards of performance, two conceptual or technical considerations are also relevant.

1. The Problem of Maximum Levels

Credentials are often restricted to those whose scholastic performance and/or test scores are higher than minimal levels required for work or other social roles. Such occurrences discriminate unfairly against those who are competent to work, for example, but who are selected out of occupational opportunities by those who believe in the simple equation: higher academic achievement means better work or life perform-

ance. The tacit assumption that superior abilities in all measured characteristics are necessary or even desirable for performance is highly questionable.[6]

Measures typically used to assess job task performance and performance relating to the mastery of units in a curriculum typically have little bearing on how subunits interact. For any given job, life task, or individual performance, component skills in one area can compensate for deficiencies in others creating a variety of combinations of individual performance levels which could theoretically add up to equivalent overall performance. Thus, minimal levels of performance on individual variables (which compromise overall competence) may have little meaning by themselves. Their interactions with respect to outcomes may have far greater significance.

We are most familiar with this problem in cognitive areas of education. We are often taught language use, verbal reasoning, spatial relationship, reading comprehension, abstract reasoning and syllogistic analysis (e.g., as measured by Miller Analogies) as discrete units of curricula. Assessment of integrated or general skills such as problem solving often do not take into account the interactive nature of skills in these subcomponent areas. Cognitive measures are used almost exclusively in assessment as if the qualities they measure did not interact, i.e., they are tested separately.

The importance of interactions, while intuitively obvious in the motor skills area, have not been carefully attended to in cognitive and social/emotional areas of assessment. Yet, once individuals have gone through a series of academic life experiences that enhance their competence in dealing with school, work, and other life experiences, the appropriate assessment task becomes that of measuring such integrated and generalized learning outcomes as the ability to cope with new problems, to find appropriate solutions, and to take correct actions.

Measures which reflect the interdependent nature of cognitive skills essential for satisfactory functioning outside of academia have only begun to be developed.[7] For example, Klemp's General Integrative Model of Assessment, incorporating a variety of independent techniques,

---

[6]A simple motor skill example will demonstrate this point: we know that an automobile driver must grip the steering wheel with enough force to maintain control of the car. But beyond a certain level of pressure, added strength in holding the wheel does not increase overall driving competency. And this is just one of some 3,400 discrete behaviors identified by researchers as making up the task of "driving."

[7]A recent example in the noncognitive area by McClelland and Burnham (1976) reports the importance of the interaction between levels of motivation and ego-maturity for managerial competence.

is an approach to summative evaluation of an individual's ability to solve a problem which has as many elements and complexities of real life situations as possible. Such an assessment of individuals has the potential of coming closer to tapping real life competence than can any single test alone.

While it makes sense to require minimal levels of proficiency for many competencies, ability levels over and above necessary cut-off points do not always correlate with overall performance.

For example, in a job analysis, McClelland and Dailey (1974) found that a minimal level or organizational or clerical competence was necessary for human service workers in the Massachusetts Civil Service system, but high scores on these measures were negatively correlated with superior job performance. Selecting people by rank according to score not only discriminated against those whose scores were adequate (sufficient) though "Uncompetitive," but the process failed to select the better job performers as well. This finding and others[8] suggest that going beyond sufficient levels of competency in awarding credentials can be very dysfunctional for society - not only in terms of equity, but in terms of meritocracy as well.

In many job situations, where cognitive and other competency measures are used to select job applicants, even if job relevance of the characteristics being tested for can be demonstrated (e.g., "verbal ability" in human service workers), level of sufficiency for competent job performances is rarely evaluated or known.

We need more empirical research to establish minimal levels of competence required for quality performance based on how workers in the field perform on various competency measures.

2. The Problem of Interactions

Researchers have long recognized that the interaction effects of variables are quite often more significant and meaningful than individual variables taken alone. It was stressed earlier that competence is not a simple summation of discretely defined skills and abilities. This is readily seen in the example of driving ability. Although one can identify many skills necessary for safe and effective driving - including attitudes, cognitive skills, and emotional factors, as well as perceptual and motor skills - it is intuitively obvious that a simple summation of measurement scores on these discrete task performances would not add up to equivalent driving skills. An individual who is overly competent at some driving skills but woefully inadequate at others would be a poorer driver than someone whose skills were all sufficient, though their summed skill scores would be identical.

---

[8]A recent study at Harvard revealed that the past SAT scores of faculty members were negatively correlated with more successful teachers (McClelland, personal communication).

The implication for higher education is that one cannot assume that abilities or skills discretely learned will be integrated in work and life functions and consequently that establishment of minimal levels of performance on isolated skills or "subcompetencies" have much meaning in themselves. Therefore, competency research, new assessment procedures, and test instruments must also focus on the interdependence of skills.

APPENDIX D

REFERENCES

American Association of Colleges for Teacher Education, Committee on Performance-Based Teacher Education. Achieving the Potential of Performance-Based Teacher Education: Recommendations. PBTE Series, No. 16. Washington, D.C.: American Association of Colleges for Teacher Education, 1974.

Atkinson, J.W. Motives in Fantasy, Action, and Society. New York: Van Nostrand, 1958.

Broudy, Harry S. A Critique of Performance-Based Teacher Education. PBTE Series, No. 4. Washington, D.C.: American Association of Colleges for Teacher Education, 1972.

Broudy, Harry S. Can research escape the dogma of behavioral objectives? School Review, 1970, 79, 43-56.

Dodl, Norman et al. Florida Catalog of Teacher Competencies. Albany, NY: Multi-State Consortium on Performance-Based Teacher Education, 1973.

Gage, N. Models for Research on Teaching. Occasional Paper No. 9. Stanford: Stanford Center for Research and Development in Teaching, School of Education, 1976.

Hodgkinson, H.L. Evaluation to improve performance. In D.W. Vermile (Ed.), Learner-center reform. Current Issues in Higher Education, 1975. San Francisco: Jossey-Bass, 1975.

Klemp, G. The Objective Measurement of Prejudice and Discrimination: Final Report to the Office of Naval Research. Boston: McBer & Co., 1975.

Klemp, G. Validation of a Broad-Based Examination of Entry-Level Support Positions in Administrative and Technical Work. Boston: McBer & Co., 1976.

Kohlberg, L. Stages of moral development as a basis for moral education. In C. Beck & E. Sullivan (Eds.), Moral Education. Toronto: University of Toronto Press, 1970.

Kolb, D.A. On Management and the Learning Process. Working Paper No. 652. Cambridge, MA: Massachusetts Institute of Technology, 1973.

Kolb, D.A. Individual Learning Styles and the Learning Process. Working Paper No. 535-71. Cambridge, MA: Massachusetts Institute of Technology, 1971.

Kolb, D.A. and Goldman, M. Toward a Typology of Learning Styles and Learning Environments. Working Paper No. 688-73. Cambridge, MA: Massachusetts Institute of Technology, 1973.

McClelland, D.C. (Ed.), The Impact of College on Personal Characteristics and Life Outcomes. Cambridge, MA: Harvard University, 1976.

McClelland, D.C. Power: The Inner Experience. New York: Irvington Press, 1975.

McClelland, D.C. Testing for competence rather than for "intelligence." American Psychologist, 1973, 28.

McClelland, D.C. Achievement motivation can be developed. Harvard Business Review, 1965, November-December.

McClelland, D.C. The Achieving Society. New York: Van Nostrand, 1961.

McClelland, D.C., Baldwin, A.L., Bronfenbrenner, U., and Strodtbeck, F.L. Talent and Society. New York: Van Nostrand, 1958.

McClelland, D.C. and Burnham, D.H. Power if the great motivator. Harvard Business Review, 1976, March-April.

McClelland, D.C. and Dailey, C.A. Professional Competencies of Human Service Workers: A Report to the College for Human Services. Boston: McBer & Co., 1974.

McClelland, D.C. and Klemp, G. Report to the Executive Office of Manpower Affairs on the Validation of a Human Service Worker Test. Boston: McBer & Co., 1974.

McClelland, D.C. and Winter, D.G. Motivating Economic Achievement. New York: The Free Press, 1971.

Medley, D.M., Soar, R.S., and Soar, R. Assessment and Research in Teacher Education: Focus on PBTE. PBTE Series, No. 17. Washington, D.C.: American Association of Colleges for Teacher Education, 1975.

Merrow, J.G. The Politics of Competence: A Review of Competency-Based Teacher Education. A report to the National Institute of Education. Washington, D.C.: Institute for Educational Leadership, 1974.

Merwin, J. Performance-Based Teacher Education: Some Measurement and Decision-Making Considerations. PBTE Series, No. 12. Washington, D.C.: Association of Colleges for Teacher Education, 1973.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-66.

Pottinger, P. and Klemp, G. Final Report to FIPSE, Contract No. OEC-0 74-9258. Boston: McBer & Co., 1976.

Pottinger, P. Comments and Guidelines for Research in Competency Identification, Definition and Measurement. No. SURC-TR 75-582. Syracuse, NY: Educational Policy Research Center, Syracuse University, 1975.

Rosenthal, R., Archer, D., Koivumaki, J.H., DiMatteo, M.R. and Rogers, P. Assessing sensitivity to nonverbal communication: The PONS test. Division 8 Newsletter, American Psychological Association, January 1974.

Stewart, A.J. Longitudinal prediction from personality to life outcomes in college educated women. Unpublished doctoral dissertation, Harvard University, 1975.

Stewart, A.J. Scoring Reactions to Rhetoric of Analysis of Argument. Unpublished manuscript, Harvard University, Department of Psychology and Social Relations, 1974.

Stewart, A.J. Scoring Systems for Stages of Psychological Development. Unpublished manuscript, Harvard University, Department of Psychology and Social Relations, 1973.

White, R.W. Motivation reconsidered: The concept of competence. Psychological Review, 1959, 66, 297-333.

Winter, D.G. Scoring the Test of Thematic Analysis for Critical Thinking. Unpublished manuscript, Wesleyan University, 1973.

CONFERENCE SESSION

Paul Pottinger
Institute for Competence Assessment
McBer and Company

Before responding to questions, I would like to sketch out some ideas important to my work.

I think one of the reasons that we have pursued the behavior objectives approach to its limits is because of the need to be able to measure learning outcomes. We need to do this for accreditation, licensing, accountability and for a variety of political, social and economic reasons. We need to be able to demonstrate that what teachers are doing is effective. This leads, of course, to the problem of how do we measure outcomes, which in turn leads to the question of how do we measure outcomes reliably--that is, how does one gather evidence that is trustworthy. How do we get evidence about a teacher's performance that is trustworthy, not only to those who have to make judgments about competence, but trustworthy to parents, accrediting agencies, faculties and to the teacher himself or herself.

When we get to this issue of trustworthiness, we are confronted with some limitations in the state of the art in assessment. Often the outcome is that we become tyrannized by the concept of reliability. For example, we frequently use what are called "low inference measures" simply because they are reliable, but this gets us into a reductionist mode of measurement. The danger is, of course, that we can reduce the operational definition of knowledge, skills and abilities to the point where they become irrelevant or meaningless, all for the sake of having "trustworthy" (i.e., reliable) evidence. Too often we end up with a tremendous amount of reliable evidence that adds up to nothing in terms of quality teaching.

One approach to resolving this problem is to try to achieve a synthesis between rigor (which allows us to get at reliability or trustworthiness) and relevance (which allows us to get at outcomes in which we're really interested). The tension between these two considerations often leaves us, on the one hand, with objective or low inference rating types of measures which intuitively are not very relevant to desired outcomes. That is, they are simplistic. An assumption in using these measures is that competence is a linear and additive phenomenon. Thus, if you get someone to complete a checklist of behaviors to some minimal level of observable performance criteria, that person is assumed to be competent. The fact is, however, if you look at any number of people who have gone through such a checklist, some teachers may still be considered more effective than others (especially when you take setting and other contextual variables into consideration).

On the other hand, if we talk about measuring qualities like warmth, empathy, flexibility, ability to work under stress and to be responsive to individual needs, we are talking about personal characteristics that are difficult to measure. Interestingly, there are thousands of measures available to assess these kinds of characteristics. So why is it that we are not using them? What's wrong? Why are we constantly reinventing the wheel or falling back on external observables? One problem is that personality psychology has dealt predominantly with nomothetic rather than idiographic data. Many personality tests are therefore good for theory development, but when it comes down to assessing an individual, most personality tests are not very effective, i.e., reliable or valid. Secondly, we understand behavior more readily than causal psychological constructs. Thirdly, we're more adept at teaching behaviors than at teaching personal characteristics. Finally, there is the problem of method variance. What we typically do when we measure personal characteristics is ask people to take a paper and pencil test that requires some kind of a forced choice, multiple choice, check the adjective, or similar type of response. These respondent test-taking behaviors are not representative of the kinds of behaviors that people use when they are operating on the environment in reality, e.g., when they teach. That is, the very method of assessing someone's personality or some personality construct is in fact inappropriate to the behaviors that people have to use when they are operating on the environment. That is one reason that assessment centers became so popular, but they are too costly and impractical for most purposes. That brings me to the final point.

If we can combine the methodological rigor of multiple choice tests with relevant outcome variables which we know intuitively to be important, then we have a powerful instrument for assessment. This combination may in fact lead to trustworthy and meaningful data that allows us to understand and predict behavior. This is the premise behind some of the measures you have just read about.

What lies behind these measures is the notion, first of all, that you must have people operate on the environment in order to assess their ability. You cannot do this with paper and pencil tests that only call for respondent behavior. It is also assumed that rating, or somehow codifying, those operations can be done in a way that is reliable so you can quantify ratings and treat them statistically. Only then can you begin to deal with the question of construct validation, predictive validation and meaning. And a final assumption is that latent content is sometimes more important in understanding job performances than the easily observed and documented manifest content. It is often the presence of specific latent content in measures which lends meaning to a test score. It is also important to note with regard to test meaning that one can have highly reliable measures without having valid measures. And more importantly, one can have reliable and valid measures which have little meaning or importance for teaching. These distinctions are critical.

The focus so far in performance based teacher education has been on the <u>behaviors</u> that we want to measure as an outcome of training.

And there is no question that where we have gone with behavior objectives is critical and basic to teacher performance. That is, many of these behaviors are necessary building blocks for successful performance, but they may not be the causal factors of successful performance. Let me point out what I mean.

In any set of people who have the ability or can exhibit a certain number of behaviors at a certain level of proficiency, there will still be differences in the way people utilize their behaviors. The presence of a behavioral repertoire which may be very clearly specified and reliably assessed does not in fact quarantee or in any way indicate that those behaviors will be effectively utilized in a particular setting, in a variety of settings or with a variety of problems. What we know is that these basic skills, which are necessary but not sufficient conditions for competence, can be demonstrated. Yet we also know intuitively that given this hypothetical set of people, who have demonstrated the necessary behavioral repertoires for teaching, that some are going to be more effective than others because they are more mature, better problem solvers, more adaptive to change in classroom circumstances, more interested in their jobs, more empathic, etc.

Once you have identified a set of behaviors, you also need to know something about some of these modifier variables or causal linking variables. If you can tap certain kinds of constructs like maturity or empathy, the chances are you can predict differences in how teachers utilize behaviors to be effective. The measures that you have just read about are aimed not at the teacher behavior level, not at the external observable level, but really at the covert, underlying, causally related level which we identify with psychological constructs. We are interested in the latent (or unobservable) as well as manifest content of behaviors. These characteristics are very important in determining whether a teacher utilizes a behavioral repertoire effectively or not.

Again, once you have developed a certain set of behavior objectives, have adequately rated them and are satisfied that you can identify whether a teacher in fact possesses the repertoire of behavior you feel is important, it is also important to include measures that define some underlying causal factors which validly determine whether that teacher can utilize these behaviors effectively. For example, we did an analysis in the U.S. Information Agency and discovered that the job description of officers stressed communication skills as an important behavioral objective. Communication behaviors can be easily taught and easily measured. What we also discovered, however, was that the superior officers had an attitude and an ability that supplemented their communication behaviors and differentiated their performances from more average or non-superior officers. These included: (1) a positive attitude about people, and (2) the ability to empathize. The point is that the superior officers were those who possessed some attitudes and abilities that were not behaviorally definable (but were measurable). I'd like to talk more about these other variables and other examples in the paraprofessional world, but I've used up most of the time. Let me entertain some questions.

Question: How would you measure empathy?

When we found the ability which we call empathy important both in the human service area and in the U.S. Information Agency, the question was how to operationalize it and measure it validly. In this case, we used Bob Rosenthal's test called the Profile of Nonverbal Sensitivity (PONS) which is a content filtered speech test, (i.e., which has removed electronically any content in the samples of speech that are being presented so that all one can cue on is the affective level). People who can understand the affective level on Rosenthal's test tend to be the same people who are judged as superior performers. Perhaps it is because they are better at hearing the "hidden agendas," so to speak, of people's spoken communication. So there is a measure that, at least, seems to have some validity in terms of identifying people who have a capability to "read between the lines," an important aspect of empathy.

Question: Should these measures supplement or supplant the list of teacher behaviors?

They should not supplant the behavioral list, but these lists are intuitively inadequate. That is, a behavioral list simply is too simplistic to tap into the complex covert variables and interactions that take place in the teaching environment. The new measures discussed in my paper do not necessarily get at the interactions either but we are trying to identify some very generic kinds of abilities that in fact are generalizable, not only across teaching but across any human service delivery occupation.

We discovered with the human service workers that the single most critical differentiating variable of whether a human service worker was competent or not was simply having faith in the ability of people to change. When we took all of the competency statements and judged behaviors according to them, it turned out that the ability to establish rapport, the ability to make home visits, to understand the laws, to understand the home environment, to understand the rules and regulations, to be able to write reports and to perform according to all the rest of the job requirements list, had little relationship to superior performance. They are important; they have to be done; and they are basic skills. If you can't write and you can't get around town and you can't establish rapport, etc., you can't be a very good human service worker. The critical thing, however, was that there were human service workers who could demonstrate these basic skills, but who were only average or below average workers. The most effective ones, the superior performers, had a positive attitude toward people, and that statistically, as well as empirically (since we could measure it) and intuitively, made sense.

Question: How does one decide what is superior performance in the first place, and then how does one know that the test predicts the ability of someone to do that?

Defining superior performance is done in a number of ways. It is frequently subjective because it gets into values. Somebody has to make

a decision. In teaching, for example, it may include supervisor ratings, pupil outcomes; it may include peer ratings in the human service area. A variety of rating systems have been used successfully. In the area of sales, for instance, in industry, it's easy to get hard data. You know that a person is selling or not selling. In teaching it is more difficult to get objective data, so you have to fall back on some kind of consensual rating system. But clearly when you isolate 10 people who have been judged superior by supervisors, and peers, their pupils are high performers, and parents are pleased, you can assume they represent a "superior" criterion group. And you can say with some confidence that there is something about that group that is different from others that you should systematically examine.

Let's assume for the moment that you have such a criterion group and that there is agreement about the criterion that set them apart as superior performers on a variety of dimensions. Assume this is well documented and the documentation is trustworthy. (We should talk more about criterion for trustworthiness, but there is not time here.) Once you have that, then the measurement issue is one of predictive validity. That is, can you select among an array of people those who are superior by using an instrument or set of instruments. Do your tests separate those who are superior from the others? If so, then you may say that your tests have some predictive validity because they differentially separate criterion groups. Technically this is called concurrent validity unless you are using a separate group from those with whom you developed your criteria in the first place.)

Once you have demonstrated that a measure does predict superior performance, then you can train to it, set performance criteria to it, select people, or evaluate performance on the basis of that measure. The reason that tests have consistently failed at separating criterion groups is because they have typically not measured the kinds of variables that superior performers manifest and/or utilize in the real world. Taking a multiple choice paper and pencil test does not reflect the types of behavior people use. They have to observe, diagnose, take in information, integrate it and synthesize it cognitively and affectively, and come to some decision, and then take some appropriate action. There is not a paper and pencil test that taps all these characteristics. The only way to tap that is to get people to observe, diagnose, integrate, synthesize and to rate their ability on these tasks. But often these are covert processes. Cognitive synthesizing is not something you can watch a teacher do in the classroom. You can observe the manifest behaviors that result from such processes, but the processes themselves are covert. If you can get at these processes indirectly by coding free operant thought patterns, for example, you can then correlate these cognitive (or affective) patterns with criterion performance of teachers. It is possible to measure such latent covert processes reliably and validly, and these measures often have more meaning in terms of competence than external observable behaviors.

The next question of course is, "Can you teach these processes?" You can teach motivation, you can teach empathy, you can teach cognitive skills and a variety of others things. Whether system X or system Y is best, needs to be empirically determined, but that's a training issue and a research issue that's beyond us at this conference.

IN RESPONSE TO THE PAPER BY PAUL POTTINGER: TECHNIQUES AND CRITERIA FOR DESIGNING AND SELECTING INSTRUMENTS AND WHAT INSTRUMENTS ARE ARE AVAILABLE

Joyce Spence
The Child Center
Springfield, Oregon

Purpose

The request to respond to Paul Pottinger's paper presents at once a grand challenge and an overwhelming task. The challenge is that of an itch to eagerly respond to an intellectual statement by one who has something important to say, in this case, to teacher trainers in special education. The overwhelming task is that of a content so wide in scope, so complex in rationale, and so detailed in length as to make choosing a starting place difficult.

The issue at hand is part of a larger concern; putting the purpose of this small paper into context of that larger concern will help orient the reader, and hopefully help orient my own direction through Paul Pottinger's paper.

The sponsors of the conference recently attended in Plattsburgh had one purpose: that was to examine several aspects of assessment of competencies, especially as those competencies relate to teachers in training in special education. It is presumed here that the reader is informed on the basic tenets of competency based teacher education. One might refer to the papers of Dr. Hugh Baird from Brigham Young University, and Dr. Gilbert Shearron, from Georgia State University, in these same proceedings, to define some of the parameters of competency based teacher training. Recently, several excellent reviews of the current literature and antecedent material have become available (see Semmel and Semmel, 1976; Houston, 1974). The direct impetus for the conference itself is a most tangible measure of the impact of competency oriented planning; that is, the State of New York mandate for demonstrated competence by graduates of state training institutions. That mandate is felt most immediately by teacher trainers in the colleges and universities. Teacher trainers in turn could be said to be suffering a sort of management dyspepsia; too much data on too many competencies taken too fast. What many teacher trainers who are working to provide a competency based program need is help in assessing the learning which has occurred, its applicability in teaching settings, its generalizability to new learning settings.

The conference sponsors had four specific objectives related to assessment of competencies: 1. to develop instruments to measure competencies other than interpersonal, 2. to develop instruments to measure interpersonal competencies, 3. to establish validity and reliability for such instruments, and 4. to develop a system to help manage the results of such assessment. Paul Pottinger's paper relates to the second objec-

tive, the task of developing ways to measure competence in interpersonal skills.

Content of Pottinger's Paper

Typically, one would expect certain rather predictable content areas in a paper whose title included key words "techniques", "criteria", "designing", and "instrument". The teacher trainer attending a conference having to do with assessment of competencies and given such a paper to read would expect an outline of types of measure, applicable under which circumstances, to be used for what order of skills, and so on. The tendency to expect such direction is related to the central problem in much competency oriented thinking. That is, the development of long lists of performance competencies generated by an expert or collection of experts which are assumed to represent the set of requisite skills in a real teaching setting. The listed competencies, in turn, are tied to some prestated criterion measure, which, when achieved, are presumed to demonstrate one's competence. The whole process, as outlined, is further presumed to follow the same conditions as one might allow for any other learning activity. An example might be the learning of a simple motor skill such as typing, or tying a shoe: task analyze; sequence; reverse chain; practice to competence; demonstrate learning of task. The result of that thinking, all too often in teacher education, is a long list of proposed competencies. Paul Pottinger states, "Most new attempts to define and measure learning outcomes according to what people can do are restricted in scope, lack of vigor or poorly correlate with job and life requirements." In his discussion of assessment issues in PBTE, Pottinger further states, "Until we have a more comprehensive base, educators will continue to use an existing array of questionable measures based on narrow cognitive outcomes, superficial behavioral observations, and a priori value-laden judgments. What is required is a sophisticated technology capable of uncovering critical indicators which are necessary and sufficient for competent performance."

The concern about surface, or superficial, relevance has been at the core of criticism from the beginning of PBTE. Perhaps more accurately, the philosophies of education represented by the structural-analytic and the unstrucured-humanist orientation as extremes have become translated in part as the competency vs. creative learning/teaching approach. A further discussion of that is not possible here except as it related directly to Paul Pottinger's primary thesis.

By setting the stage for the problems currently experienced in PBTE, especially the limited value and narrowness of many of the competencies, Dr. Pottinger leads us to "locating a 'rational center'", or "moving from behaviors to their causes". The continued direct quotation here is probably superior to anything I could write. "...teacher competency definitions and assessment instruments should compromise causally related, criterion referenced and generalizable variables which take into account complex observable and unobservable behavioral repertoires required for successful performances." The key words here are causally related, complex, and unobservable (underlining mine). The narrowness and limited use of most PBTE assessment techniques and the competencies they are related to is because they are "reductionistic and lacking in meaning (and) because they fail to assess the underlying causes of these behaviors" (un-

derlining Pottinger's). Pottinger concedes that such a point of view will generate strong reactions from PBTE supporters, and, in my opinion, no less from the opponents of the medical "cause-oriented" point of view in general.

Implications for Special Education

The area of education designated "special" has had a relatively short and stormy history. Central to much of the storm are the efforts to define what is special about "special". At one time the population of children which received special attention were used to define "special". That categorization continues to some extent today although the limitations of categorically defining children as being special because they receive special help begins to reach the outer ranges of tautology. "Special" has also been defined in terms of the scientific basis of its application. "Special" educators often prefer to identify themselves as an applied psychology rather than a branch of education. "Special" has sometimes included special equipment, materials or gadgetry. Recently, back to the PBTE topic, "special" has meant many additional diagnostic and remedial competency requirements over regular elementary education, to meet special education certification requirements.

A perspective on "special" which I have not seen discussed is the area of survival competencies. By survival I mean the ability: 1. to tolerate ambiguity, 2. to prioritize problems and deal with them in order of importance for the child, 3. to maintain oneself on a lean reinforcement schedule from your clients (students) or to see and appreciate very small and infrequent gains, 4. to tolerate verbal or psychological abuse from disturbed children, 5. to recognize ones own personal and interpersonal needs and assertively work to meet them. My twelve years in special education have taught me that these are the important competencies in "special" education and the things one must learn to survive with handicapped children.

In Relation to Interpersonal Competencies and their Assessment

The purpose of interjecting my personal definitions of some aspects of "special" is (1) to demonstrate my conceptual agreement with Paul Pottinger that being competent must ultimately relate to some basic problem-solving, learning-to-learn, wholistic perspective, and (2) to relate that notion of "special" as a personality construct which encompasses interpersonal skills as a part of the whole.

The application of that thinking to a teaching training program would have some obvious results. One would be a learning by doing, or field-based orientation. Most programs involve field settings. Another would be to develop divergent thinking through course work and actual problem solving. Some programs do that. And finally, such a perspective on self-awareness and the relation of a job performance to that awareness would result in training and education experiences to tap those unobservable and causally-related dynamics. Few teacher training programs that I know of do that.

Any discussion of teacher training methodology must first accept that adults learn exactly the same way as children. We know how to teach chil-

dren academics. We are learning how to teach children self-awareness, acceptance of differences, and other interpersonal skills. Student outcome for teachers in training would then necessarily follow the same patterns as pupil outcome for children in class. In relation to interpersonal skills we are concerned with the fruition or melding of cognitive and affective dimensions of competence.

Common goals for both which are difficult to behaviorally observe and measure are:

1. ability to communicate effectively under stress;
2. knowledge of one's feelings and the concomitant physical sensations of emotion;
3. a realistic perception of one's self and the effects of one's behavior on others;
4. the ability to behave flexibly in relation to one's inner state and accept one's own inconsistencies;
5. the ability to be empathic, or to "be with" another person emotionally.

These presumed abilities are examples of what is probably a far more complex psychological construct than we can imagine. We do know that they are at least some of the abilities requisite to reciprocal positive interactions with others, so-called "healthy" functioning inter-psychically, and positive cultural attributes.

Further, we can specify some behaviorally observable objectives which provide material for training and assessment. Examples of some of those are:

1. paraphrasing to assure understanding of what others say;
2. describing behavior, rather than interpreting;
3. describing feelings, rather than expressing feelings;
4. checking perceptions of other's feelings or intentions;
5. identifying nonverbal communication cues;
6. giving and receiving feedback;
7. identifying the effects of expectations in communication;
8. applying the concept of matching behavior with intentions in communicating;
9. identifying the influence of one's personal style of communicating;
10. applying techniques of assessing one's knowledge and skills in interpersonal communication;
11. applying guidelines to create self-improvement in communication.

To meet these objectives in a wholistic problem-solving manner requires the student to learn to communicate by communicating. The interaction itself is the behavior from which data is collected. The interaction is structured around a very specific objective and then provided with the life material of the persons involved. Specification of the problem, data collection thorough observation, structured feedback, and guided summarization are built into each exercise for each objective to meet each goal. Students must problem solve to learn problem solving. They are required to communicate under stress and receive feedback. They work to learn to "be with", or be empathic with another person through paraphrasing, feedback, and perception checking.

How is this evaluated or how do we know when the students are competent in interpersonal skills? Several ways provide data, although none yield a

percentage test score. The most important way is to ask the persons themselves, or allow them to ask themselves. Provide criteria for them, and require them to respond in life-problem situations. Another way is to require the student to accept a "learner" style. That is, require the student to develop certain skills, collect and maintain data on that over a period of time outside class, demonstrate growth in the behavioral diary, and perform in group in class. Like children in reading, we are not as concerned with meeting certain criteria set for everyone, but instead to demonstrate a steady growth and reinforce a positive attitude toward learning. And finally, we ask the members of the group to be constantly helping each other to grow by reinforcing the skill of providing feedback and the notion of growth and behavior change through input from our world and practicing new behaviors.

In the area of interpersonal skills, maybe more than in any other teaching skill, the role of the trainer in evaluation is complex and problematic. Ascribing certain cultural expectations on interpersonal performance can be ineffective if not downright risky. Rather we must look to the individuals themselves to develop some skills in remediating their own interpersonal behavior, in eliciting positively reinforcing feedback from those around, and monitoring their own internal state of equilibrium. We must also take a collective responsibility for making effective interpersonal behavior an important criterion for teaching, and for taking a personal responsibility to make it a primary criterion for living.

My thanks to Dr. Paul Pottinger for writing an important paper for teacher educators. I believe the direction of his thinking will result in creative changes in the current focus of training activities.

REFERENCES

Semmel, MI.I and Semmel, D.S. Competency-based teacher education: An overview. Behavior Disorders, 1976, 1, 69-82.

Houston, W.R. (Ed.), Competency Assessment, Research, and Evaluation. Washington, D.C.: American Association of Colleges for Teacher Education, 1974.

SPECIFYING WAYS AND MEANS OF INSURING RELIABILITY AND VALIDITY IN COMPETENCY ASSESSMENT

Beatrice A. Ward and William J. Tikunoff
Far West Laboratory For Educational Research and Development

As we understand our task here, we are to offer for your consideration a field based model for validating teaching competencies, particularly as they relate to teaching the handicapped. Within the parameters of this assignment, it is important first to examine the current states-of-the-art both of educational research and the competency-based teacher education movement, and then to define those operational variables necessary to assessing teaching competency. Second, the issues of establishing reliability, validity and generalizability of assessment procedures must be addressed. Finally, a description of a process must be offered for accomplishing context-based assessment of and training for teaching. This paper is organized to address each of these concerns in turn.

RATHER CONTEXT-BASED THAN COMPETENCY-BASED

As can be perceived from the title of this section, we will argue that to date any attempt to assess a teacher competency has been a single variable procedure, and that what is needed instead is a multivariate approach. This is to say that the CBTE movement has tended to treat teacher competencies in a generic fashion, assuming that such competencies can be applied across all teaching situations and fit comfortably into each. Thus, when assessing such competencies, the tendency has been to look at the performance of the teacher as the most important, single evidence of attainment of a competency. Further, if generalizability of competencies has been attended to at all, it has been assumed that once one had accomplished attainment of a competency, one easily could perform in a similar fashion whatever the setting.

It is precisely this stance which we feel is unrealistic and with which we take issue here. The act of teaching cannot be investigated successfully unless one takes into account the context in which it occurs. By this we mean that a single teaching competency cannot be considered in a vacuum. We must also take into consideration the total context of the teaching/learning event in which the teaching competency occurs. It is this view of teaching which Smith (1976) labels "multimethod, multiperson, multisituation, multivariable".

In order to view teaching in this way, then, one must attend to the myriad of variables simultaneously in operation. For instance, although focusing on a single teaching competency, some context variables one might consider are:

(1) Student population variables: What grade level, age level? How many? What percentage are handicapped (physically, mentally, emotionally)? How seriously? What level of SES? How many are habitually

absent? How many are recently enrolled in class? How assigned to teacher (homo- or hetero-geneously)? How many speak at home a language other than English? What is the boy-girl ratio? How many had breakfast before coming to school? How many like school/the teacher? How many cannot read? Are students engaged productively in the teaching/learning activity being observed? Who/how many are receiving instruction outside of school/what kinds?

(2) Curriculum variables: What is the content/subject area in which the teacher will be observed? Is the curriculum imposed or is it teacher-originated (Another way of asking this is,"To what degree was the teacher involved in creating the curriculum?")? Do the curriculum concepts to be attained "fit" the students developmentally? Is it a curriculum the students enjoy? Is it a curriculum that parents support? Is it consistent with current knowledge, i.e., consistent with their values?

(3) Instructional materials variables: Is there a match between curricular purposes and instructional materials? Whose materials, teachers' or publishers' If publishers', who selected them (did the teacher have a voice in selecting materials)? Are there enough materials so that each student has access to them? Are materials at a realistic reading developmental level so that they are comprehensible to students? Does the teacher like the materials? Do the students like the materials?

(4) Classroom setting variables: Is the classroom comfortable, i.e., well-lighted, bright, cheerful--in short, a place where you would like to be? Are there enough places to sit so that each student can work comfortably? Is there enough space to accommodate everyone without being too crowded? How many blackboards? bulletin boards? bookshelves? closets? Is there more than one adult in the classroom, e.g., teacher aid, student teacher, other teacher, etc.?

(5) Organization for instruction variables: Is the organization for instruction consistent with curricular goals? Are the students organized for instruction into groups? If so, how organized? What criteria were used, e.g., alphabetically, I.Q., reading scores, etc.? Do students work independently? If so, where, e.g., at learning stations, at own desks, outside the classroom, etc.? Are students taught as a whole group? If so, how are students assigned to desks? Who sits nearest the focus of instruction?

(6) Allocation and utilization of time variables: How much time is allocated for the teaching/learning activity under observation? How much in relation to other activities? How does teacher spend his/her time, e.g., with which students, on what tasks, instructing or controlling behavior, etc.? When is event occuring, e.g., first thing in the morning, before/after lunch/recess, last thing in afternoon, at beginning/end of year, before/after a holiday, etc.? How do students spend their time, e.g., who is engaged in assigned activity, and who is not engaged?

(7) Distractions from instruction opportunity variables: Are all the students present, or are some absent? If absent, are they

frequent absentees? How many students are recent transfers into the class? Is teacher allowed to teach the entire time, or are there interruptions, e.g., notes from office, fire drill, unexpected visitors, etc.,? What is the noise level? What is known about individual students' tolerance for distractions?

(8) Teacher variables: How secure does the teacher feel in teaching the curriculum under observation, e.g., does (s)he feel (s)he "knows" the subject matter? Is it a favorite subject? Would (s)he rather be teaching something else? Is this the first time (s)he has taught this curriculum, or is it a familiar one? How many years has the teacher taught, e.g., student teacher, new teacher, taught 1-3 years, taught 3 years or more? Is the teacher alone in the classroom, or part of a team? What is the teacher's most comfortable"style", i.e., whole group, small group, individual instruction, and how does this "fit" with the prescribed competency? Is the teacher in good physical/mental health? What are his/her concerns about education? Does (s)he like the students, and believe that the students like/do not like him/her? Does (s)he like the school? the principal? other faculty? Is (s)he at this school by choice? How secure is his or her job with the system and in this school?

What is intended here is merely a partial listing of some context variables in order to illustrate those dimensions of a teaching/learning event which must be taken into consideration if one is to assess the performance of a teacher in relation to a given competency. Later in this paper, we will demonstrate the use of such context variables in situ and in relationship to the issues of reliability, validity, and generalizability.

It is clear that such a multivariate view of teaching differs markedly from the single variable approach, but it is not without precedence. The Spring 1976 Journal of Teacher Education -- focusing on a conference held in Austin, Texas in October, 1975, at which those engaged in research on teaching effectiveness reported their current findings -- is most revealing in this respect. Our own contention (Ward and Tikunoff, 1975) that experimental and quasi-experimental research methodologies which attempt to isolate and manipulate independent variables in a controlled setting are not as productive as naturalistic studies which allow events to occur realistically, is supported by the work of Brophy (1975), McDonald (1975), Stallings (1975), and Tikunoff, Berliner and Rist (1975) cited in that Journal. Further, Berliner (1975) cites fourteen impediments to the study of teaching, all of which bear upon and significantly support our thinking here.

Given that research on teaching effectiveness is beginning to attend to the multiplicity of factors that comprise a teaching/learning event, it seems appropriate to develop a multivariate procedure for assessing the performances of the act of teaching.* We turn now to this task.

---

* We note with some alacrity that both research on teaching effectiveness and the CBTE movement have involved the classroom teacher only at that point when consistent advocacy is paramount to survival. We wonder: If practicing classroom teachers had been involved at the inception of most educational research and the CBTE movement, would it have taken us as long to reach our conclusions? Later in this paper we propose such parity.

A Multivariate Approach to Assessing Teaching Competency

In our years of learning, teaching, training teachers, and researching teaching effectiveness** we have been fascinated with the ease and precision with which some teachers can accomplish a teaching act which proves to be very difficult and most uncomfortable for other teachers. Conversely, given a different teaching act, the situation might be reversed and those who did poorly with the first do well with the second. How is it, then, we ask, that as judges of "good teaching" we insist that all teachers strive to reach something we call the maximum level? And with the same skills?

In the real world, such is not the case. We offer ourselves as examples. Both of use have taught the spectrum of school curriculum, and in this respect Betty was outstanding as a math teacher and William as a teacher of writing. Both of us excelled at teaching reading. We both like to cook, but Betty eats right-handed and William eats left-handed. We enjoy good music and both are full season subscribers to the San Francisco Opera; however, Betty plays a musical instrument, while William does not. The point we wish to make here is that in the real world individual differences abound, but they do not impede people who have disparate interests, skills, experiences, and personalities from reaching common objectives, each in his/her own way.

In the real world, the attainment of an objective by different people allows for a variety of ways of getting there. In addition we recognize, given any task, that different people most probably will perform at different levels of expertise, and that their level of performance most probably will have a great deal to do with and grow out of their interest, personality, and experience in relation to that task. To borrow from our example, even if William takes lessons and practices long hours, he probably will never play the piano as well as Betty can. However, both of us accede to the proposition that a Rubenstein Betty is not! Finally, some people are simply better at one thing that at another, and each of us makes a whole array of life's decisions unconsciously based on knowing what-it-is we are "good at" and what-it-is we are not.

As teachers, we understand the individual strengths and abilities of students, and we accommodate these accordingly. We recognize that all of our students probably will not attain grades of "A" for a given subject, and this does not distress us particularly. In fact, we are not certain that we wish it to be possible, because we recognize the inherent value of individual differences.

As trainers of teachers or as judges of "good teaching", however, we seem to negate our good sense and insist on the impossible: that to be successful a teacher must be all things to all students. With this

---

** We totalled these once. As a frame of reference, it is of interest to note that collectively we represent 22 years of teaching from grades K-12 and adults, and 20 years of developing training/training teachers and researching teaching effectiveness.

as its tenet, it seems to us, the CBTE movement has set about the task of compartmentalizing teaching/learning into discreet acts called "competencies", assuming in the process that all teachers can--or ought to be able to--achieve each of these. We apparently have neglected to apply our own innate humanity to assessing teacher performance in the same ways we claim to assess the performance of children in the classroom.

To make some sense out of this Sisyphaen task of assessing teaching performance within the frame of Context-Based Teacher Education, then, we have suggested here that four conditions must be accommodated. First, we believe that the act of teaching occurs not in a vacuum, but as one component of a holistic classroom ecology composed of many other components which here we have called context variables. Second, given a common instructional objective for the same student we believe it to be possible that two teachers with contrasting styles can effect similar results. This is supported by Merton's (1957) notion of functional equivalents in sociological theory in which he points out that, within a social unit, quite different overt manifest actions can lead to the same consequences. Third, depending upon teachers' experiences, training, personalities, and interests, a given teaching act or competency will be more or less difficult, and this in turn will be affected by these context variables. Smith (1976), for instance, states that "As teachers hold different views...different events will occur in the classroom and the teaching will be subject to different judgments of effectiveness". Fourth, if the three conditions stated above are acceded to, then a teaching competency realistically can be assessed only by using a context-based approach, attending to the multimethod, multiperson, multisituation, multivariate aspects present and bearing upon the performance of that competency.

Within these parameters, a given teaching competency must be viewed bifurcately. First, one must look at the teaching competency to be assessed in light of the context variables which are simultaneously in operation. Second, one must assess the performance of the teacher.

In the first instance, one is attempting to describe the complexity of the total instructional ecology in order to establish the degree of difficulty of performing the given teaching competency within the context of that ecology. Obviously, the degree of difficulty may vary among teaching situations. For example, consider a competency like individualizing instruction and conducting individual activities, listed as Competency 34 in Resources for Performance-Based Education (Houston, et.al, 1973). Such a competency can be assigned a degree of difficulty ranging along a continuum from "very easy" to "very difficult" depending upon the degree to which each of the context variables impinges upon and makes more complex the performance of that teaching competency. In essence, it is much easier to individualize instruction given a classroom of 20 suburban, high-achieveing, self-directed students than it is in a classroom of 35 inner-urban, non-reading, hostile students.*

* We have intentionally kept this example simple. One needs to assess each situation in light of the remaining context variables presented earlier in this paper to establish a more complex and holistic description of each of the two classrooms presented here for contrast. Depending on such description, the competency cited here may grow easier, become more difficult, or remain the same.

In the second instance, the level of performance of a teaching competency may be assessed by observig the performance of the teacher. Of necessity, this is an active judgment in that it takes place while the teacher is performing the teaching competency. Such an assessment necessarily must take into account the experiences, training, belief systems, and personalities of the judges; and these serve to detract from making an objective judgment. For this reason, we maintain that viewing the performance of a teaching competency by several people provides more information than does a single viewer, and we assume this stance in the final section of this paper which describes a way to put this process into operation.

The paradigm for assessing teaching competency within Context-Based Teacher Education which we are proposing here has its origin in a metaphor proposed by Lou Smith.* He was reminded while watching springboard and platform diving competition that both the degree of difficulty and the level of performance were considered in judging individual dives. In competition, the degree of difficulty of a dive is set and agreed upon prior to the performance of a single diver. This preactive judgment is no doubt the result of much negotiation among coaches, judges, divers, and the governing boards of the various athletic organizations which sponsor such competition. The judges then witness the dive as performed by an individual diver and assign a number score of from 0-10. This level of performance score is derived by rating the performance of the diver along aesthetic principles such as angle of the body upon entering the water, "perpendicularness" to the water at crucial junctures of the dive, whether the diver "goes over" when (s)he enters the water, how much splash (s)he effects upon entering the water, etc. Also, the performance always is subject to influences of personal aesthetics, such as whether or not a judge liked the dive as a total effort. In order to ascertain a total score for an individual dive, several judges judge the level of performance and this is multiplied by the degree of difficulty of the dive, with the highest and the lowest scores thrown out and the rest added up.

It is such a process that we propose for assessing a given teaching competency. By attending both to the degree of difficulty within the context of the total instructional ecology as well as the level of performance, we believe that the act of assessing competency is placed in a better juxtaposition to the reality of the classroom.

To summarize, what we have done here is to define those operational variables necessary to assessing teaching competency. Having identified such variables and introduced an assessment paradigm for operationalizing them, we turn now to a discussion of the procedures for establishing reliability, validity, and generalizability. Finally, we will propose a process for accomplishing context-based teaching which can be shown to be reliable, valid, and generalizable.

---

* Louis M. Smith, "Effective Teaching: A Qualitative Inquiry in Aesthetic Education," a paper presented at the Workshop Exploring Qualitative/Quantitative Research Methodologies in Education, San Francisco: Far West Laboratory for Educational Research and Development, 1976.

CONTEXT-BASED RELIABILITY, VALIDITY, GENERALIZABILITY

Further complexity is added to what we have proposed here when we view competency as being context-based. Each competency--if such a thing as an isolated competency exists--must be considered as an individual matrix, the component parts of which are multimethod, multiperson, multisituation, and multivariable (Smith, 1976).

To illustrate, let's take a competency labeled organizing for instruction. From the multimethod dimension this may include organizing to instruct either the total class, or small groups which include students with similar learning needs and abilities, or small groups which include students with differing needs and learning abilities, or individuals within the classroom, or individuals in special programs outside the classroom, or various combinations of small groups and individuals within and outside the classroom. Any one of these methods may be used.

To continue, the multiperson dimension of the competency might include:

- the students--What types of teaching/learning experiences have they worked in previously? Is the purpose of this experience to carry out instruction as efficiently as possible? To introduce the students to a new teaching/learning experience? Do the students have similar learning needs? Do some students have special learning needs? How independent are they?

- the teacher--Has the teacher had training in some, but not other, forms of organizing for instruction? Is one method more comfortable than another?

- other persons such as a special teachers, paraprofessionals, the school principal, parents--Are the services of one or more of these persons available? What forms of organization best use their services? Is one form of organization preferred by the school principal? The parents?

When the multisituation and multivariable dimensions are added to those of method and persons, competency becomes even more complex. Staying within the confines of the classroom, multisituations might include using/not using textbooks, using/not using audio-visual materials, using/not using computer terminals. It might include a large or a very small amount of time set aside for instruction. Limiting the multivariable dimension of the competency to student learning, one must consider whether the students are to acquire facts, skills of logic and analysis of scientific and/or artistic meaning, skills of cooperating with others to achieve a common goal, etc.

Needless to say, the matrix for this single competency is very involved. While the example we have provided is incomplete, it is sufficient to illustrate our concerns regarding exactly what about teaching

can be studied from the perspectives of reliability, validity, and generalizability, each of which purports to test a given teaching competency within the parameters of the quantitative psychostatistical research methodology. It is the traditional application of these three tests which concerns us here precisely because they take on new dimensions when dealing with a teaching competency within a multimethod, multiperson, multisituation, multivariable framework.

Patton (1975) suggests that reliability is concerned with the replicability and consistency of scientific findings. Deutscher (1970) presents a similar view but he suggests that we need to be concerned with what it is we are being consistent about. He asks whether, when we rely on traditional tests for reliability, we instead are learning a great deal about how to pursue an incorrect course with a great deal of precision.

To carry this definition into a classroom context, can you as a teacher consistently perform a competency in the same way, selecting the same method, person, situation, variable combination each time? And do you use the competency at the same level of skill each time, never improving your teaching? If so, you are reliable. But are you effective?

While this question may be so ridiculous that you wonder why we are presenting it, we would like to remind you that all of us have spent a great deal of time trying to be consistent in the inconsistent world of teaching and learning.

Patton (1975) goes on to define validity as being concerned with the meanings and meaningfulness of the data collected and the instrumentation employed. He gives particular emphasis to the meaning of the data for the participants, in our instance the students, the teachers, and the teacher trainers. Are we observing, measuring, assessing, judging what we think we are?

Given that meaning is an important aspect of validity, we can assume that value judgments will enter into our deliberations. We can no longer be content that we counted an event, or an act, or a behavior. We also must determine whether that count represents a "good" or a "bad." And this "good" or "bad" may be based upon a variety of value systems, e.g., do you value academic accomplishment, socialization, artistic development, experiencing as an end in itself? Therefore, we must be concerned that the event we count is identified and interpreted in a similar way by the observer, the teacher, and the student. We suggest that few efforts to assess the validity of teaching competencies have taken all these views into account.

A Positive Resolution of the Problem

Having presented you with a major dilemma--a view of teaching that attends to levels of difficulty and levels of performance within a multimethod, multiperson, multisituation, multivariable context--it is apparent that we are encouraging you to extend your thinking from a component, competency-by-competency analysis of teaching to a holistic analysis. A holistic analysis attends to the entire reality in which teaching occurs. It considers not only the quantitative occurrence of

a teaching competency (it was observed and counted) but also the qualitative influence upon all of those involved and the implications of the competency within each particular occurrence.

Research on teaching has just recently begun to stress the combination of quantitative and qualitative methodologies in holistic analyses.* Building from our work, we find that within such a data collection/analysis/interpretation framework, reliability and validity assume a different meaning. That aspect of teaching which can be judged to be reliable and valid is the appropriateness of the teaching act given the context in which it occurred. By appropriate teaching, we mean teaching that:

- takes into account the multimethod, multisituation, multiperson, multivariable posibilities that exist in a given teaching/learning context,
- is responsive to these multiple factors, and
- establishes, as a result, a set of conditions that maximize the potential for the teaching/learning event(s) to be successful, e.g., achieves the desired goals for the teacher and the students.

As you probably already have anticipated, the study of and training for appropriate teaching requires that we expand our various methodologies far beyond measuring in isolation what we in this paper identify as components of teaching, e.g., single teaching skills. It requires that consideration be given to the vast array of paradigmatic perspectives that make up the context of teaching and learning.

Reliability then becomes a measure of the extent to which teaching consistently utilizes appropriate teaching/learning conditions and carries out appropriate interactions and other acts and strategies. In other words, is the teaching consistently appropriate given the context in which it occurs?

Validity then becomes a multidimensional judgment of the appropriateness of the teaching that is observed/measured/assessed. Further, within such an approach, validity of teaching must be established before reliability can be considered. And if it is to have meaning, validity should be based upon the judgment of the teacher and the student as well as the researcher, observer, and trainer.

Establishing Validity

Validity can be established for various segments of the teaching/learning process, e.g., a complete unit of instruction which may encom-

* Our own work in progress, "Study A: A Case Study of Teacher Allocation and Utilization of Teaching/Learning Time in Mathematics During Regular Classroom Instruction," and "Study B: A Case Study of Teacher Allocation and Utilization of Teaching/Learning Time During the Teaching of an Experimental Teaching Unit in Mathematics" provide examples of such recent efforts. The work of Louis M. Smith at Washington University, St. Louis, Mo., is another example.

pass several days or weeks of teaching/learning; a single instructional period; a portion of an instructional period. What cannot be done--and this is important because it differentiates between the holistic and component approaches to the study of teaching--is to judge the segment as a separate unit complete within itself with no present, prior, or future context.

Our own work illustrates this point. Using certain of the multimethod, multisituation, multiperson, multivariable dimensions outlined earlier, namely:

- a specific curricular content, in this case a pre-stated objective in a mathematics unit on measurement;
- a set of manipulative teaching/learning materials to accompany the unit;
- a method of organizing for instruction selected by the teacher, in this case a small group;
- a group of students assigned to the teacher by the school district, in this case average to below average 9-10 years old, many of whom are Spanish speaking;
- A teacher with extensive training in the use of the inquiry process in mathematics instruction;

we ask the teacher and an observer to identify those points during the instructional process for the entire unit where appropriate and inappropriate teaching occurred. The following teaching event was identified as inappropriate by the observer, and appropriate by the teacher.

> The teacher now turns around and gets a checked box. It is about the size of a shoe box. She puts it down on the ground and she asks the students "Can you measure that?" The students all gather closer around the box. They reply, but not in unison, "yes." The teacher now turns around and takes out a string. She asks the students "What can we measure about this box?" The teacher then calls on James. James picks up the string and stretches it out but he does not say anything. He keeps looking at the string. Apparently James does not know the answer. The teacher then asks him "Can you measure the height?" A girl now comes into the classroom. Her name is Theresa. She sits down in the circle. The teacher asks the students to tell Theresa what they have done so far.*

The observer labelled this an example of inappropriate teaching because the teacher did not allow James an opportunity to respond to her question before asking for review. The teacher felt it was appropriate teaching.

---

* Taken from protocols of context-based observation of teaching/learning, collected for our work cited earlier.

She felt the time taken to review for Theresa what the group had done would give James an opportunity to figure out the answer without being embarrassed in front of his peers. Since James had had many failures in the past, moving to the review had seemed more appropriate than forcing James to give an incorrect answer.

Appropriateness in this instance, therefore, considered past performance of the student, and future developments in the lessons as well as the particular teaching event under consideration. While we have yet to develop a procedure for collecting such data, it also would be interesting to know how the student responded to and felt about this event in order to judge it more accurately as an example of appropriate teaching.

To provide an additional example of the importance of using multiple inputs to establish validity of a teaching event let us consider an event that occurred two weeks later in the same classroom.

> The teacher walks by Steven. She is looking at his paper. She asks Steven if he knows the perimeter. She then explains to him what perimeter is. The teacher is now checking a boy with glasses. She tells him to go to the teacher assistant. (This is the first day in the 3 weeks that this boy has been in the classroom.) The teacher is checking Anthony's work. She now goes to Selma and checks her work. The teacher goes back to Steven. She is looking at his paper and asks him, "How would you find the perimeter?" Steven does not know. The teacher helps him.

Both the teacher and the observer considered this an example of appropriate teaching. The teacher followed up to be sure Steven understood her earlier explanation. Would you agree? Or was the teaching act inappropriate inasmuch as Steven did not learn how to measure perimeter?

You would need much more information about such factors as the student's earlier experiences with the concept of perimeter, his learning style and rate, the type of learning task in which the student was involved, the details of the explanation given by the teacher, before making this judgment. A judgment based on a single perspective may or may not be correct. With teacher and observer independently identifying the event as appropriate, we have more trust in the assessment than we otherwise would. A trainer's and researcher's perspective also would be helpful.

To extend this latter example to another step in the measurement process, both events might be judged appropriate but the first might be marked at a higher level of appropriate teaching than the second, once all factors that impinged upon the event had been considered. In other words, the teacher's level of performance might be higher in one situation than another. Therefore, we have a validation process that involves teacher, observer, trainer, and researcher judgment of the appropriateness of a teaching event. And we have the potential for judging the level of performance of teaching according to not just an appropriate/inappropriate judgment, which must be made first, but also according to a scale that accommodates various degrees of appropriateness.

At this point, we consider it essential to re-emphasize that decisions regarding appropriateness, and thus validity, must take into account the multimethod, multiperson, multisituation, multivariable factors discussed throughout this paper. Further, although we have underlined the importance of obtaining multiple perspectives regarding appropriateness, when establishing the validity of a context-based teaching event, the teacher's views as a participant observer are particularly important. In our opinion, assessment of teaching must build from an inside/outside perspective. This becomes even more critical when one is working with students with handicaps of various types. Neither the teacher nor a single non-participant (outside) observer may have the knowledge and insights to judge as appropriate a particular teaching event. Multiple views may be required to establish the appropriateness and the level of performance of that event.

Establishing Reliability

Having determined that a teaching event is valid--that is, it has been observed and judged from multiple perspectives and these judges agree that it was an appropriate teaching act given the context in which it occurred--we can move on to consider the issue of reliability in teaching.

In our view, the aspect of teaching about which it is important to be consistent (reliable) is appropriate teaching. A teacher who consistently carries out appropriate teaching acts will be an effective teacher.

The questions to be addressed relative to reliability then are:

- How reliable is the performance of a particular teacher? This might be called inter-event reliability for that teacher.

- How reliable is the use of a particular competency given that it occurs in contexts that have similar multimethod, multisituation, multiperson, multivariable dimensions? This may be inter-teacher as well as inter-event reliability.

The first question allows us to consider the reliability of very different teaching events inasmuch as judgments of appropriateness are the data base to be used. Within this framework, however, it may be necessary to establish one inter-event reliability for teaching events with a low level of difficulty and another for those with a high level of difficulty. When a teacher is attempting to grow and improve, a high reliability rating should be obtained for those competencies that are used within a familiar context matrix. A lower reliability would be expected whenever one or more "new" elements have been added to the matrix. In fact, we would hope for low reliability in this latter instance since the teacher would want to utilize a new strategy more appropriately over time.

Moving to the reliability of inter-teacher use of a given competency, the entire multimethod, multiperson, multisituation, multivariable competency matrix must be considered. We cannot establish the reliability of a procedure for assessing/observing teaching and/or the reliability of teaching performance across two or more teachers without first knowing whether the contexts and thus the levels of difficulty and appropriateness ratings have a reasonably common base.

This leads us to another concern--the functional equivalence of two or more appropriate teaching events. Returning to the organizing for instruction competency, our findings regarding teacher allocation and use of time during mathematics instruction provide an example.

The quantitative data we obtained include stop-watch timing of the seconds used for various instructional activities. In two classrooms we found the following uses of time. In a 22 minute mathematics period, Teacher A devoted 11.08 minutes to teaching that was related to math instruction. All this time was spent with the total class. Fifty-eight seconds were devoted to behavior control: 17 with individual students, 23 with small groups of students, 18 with the total class. The remaining time was spent in non-mathematically oriented activities. Teacher B, in a 23 minute period, devoted 17.35 minutes to teaching related to instruction: 15.31 minutes were used in working with individual students, 2.03 minutes with a small group of students. No time in Teacher B's class was used for behavior control. The remaining time in this class also was used for non-math purposes.

We have not as yet completed the summarization of the daily work outcomes* for the students in these classes on these particular days. However, for purposes of illustration, let's assume both groups of students performed equally well in mathematics on those days. These two very different teaching performances then could be judged to be functionally equivalent.

The next question to be addressed would be whether one teacher's way of organizing was more or less appropriate than the other. To answer this, we would need to consider the multiple factors in the context-based matrix as well as the time counts and student outcomes. Having done so, we might assign these two teachers who used a competency in very different ways a similar performance rating and thus establish a high degree of inter-teacher consistency.

Such a view of reliability differs considerably from the traditional approach which would have found little, if any, basis upon which to say the performance of these two teachers was similar, let alone, consistent.

To reiterate, whether you are concerned with the reliability of an assessment/observation/judgment procedure and/or the reliability of teaching performance, both context dimensions and functional equivalence must be considered. To do this, a holistic approach must be taken to establishing reliability just as it was to establishing validity.

Where Are We In Relation to Generalizability?

There is a major difficulty with the approach to the study of teaching we have proposed; the data base for the inquiry is an N of one.

* We are not taking the time to develop this student performance measure in this paper. It is part of our current work, and is an approach that combines the level of difficulty and level of performance principles in assessment of students. It uses daily work rather than an achievement test as the data base.

How, then, can we expect to establish a basis for prediction and generalization from one N to another N?

The answer appears to be one of identifying those multimethod, multiperson, multisituation, multivariable factors about which we collect information with each and every N. Having done so, we may be able to determine when one context-based event looks/does not look like another; and we, therefore, can/cannot expect to generalize across the event's.*

A national assessment of teaching under these guidelines would be built from a very different theoretical base than one utilizing a component approach. In a component approach, we may expect to study the same teaching across all sites. Here we are saying the context, and therefore the teaching, may be entirely different in each site that is studied. Our problem as researchers and trainers is to recognize when sites are very similar as well as very different and to use this knowledge to help us interpret and apply the information we obtain as we explore teaching site by site.

A PROCEDURE FOR ASSESSING/DEVELOPING TRAINING FOR CONTEXT-BASED TEACHING

In this paper we have argued for a view of teaching which is multimethod, multiperson, multisituation, and multivariable. While we recognize that such a view is in apposition to the traditional, single variable approach, we have shown that it more precisely accommodates the psychostatistical concerns of reliability, validity, and generalizability given the reality of the classroom. Throughout this discussion, we also have insisted that procedures for assessing teaching competencies within this paradigm--if they are to identify appropriate teaching--must attend to and include the perspectives of all those affected. It is this multiperson approach which we believe is most promising in studying and improving teaching.

In our own work in researching teaching effectiveness, we have proposed and are in the stages of planning to test an interactive research and development model which is multiperson in nature (Ward and Tikunoff, 1976). We believe it has implications as a vehicle for accomplishing the tasks and procedures we have advanced here.

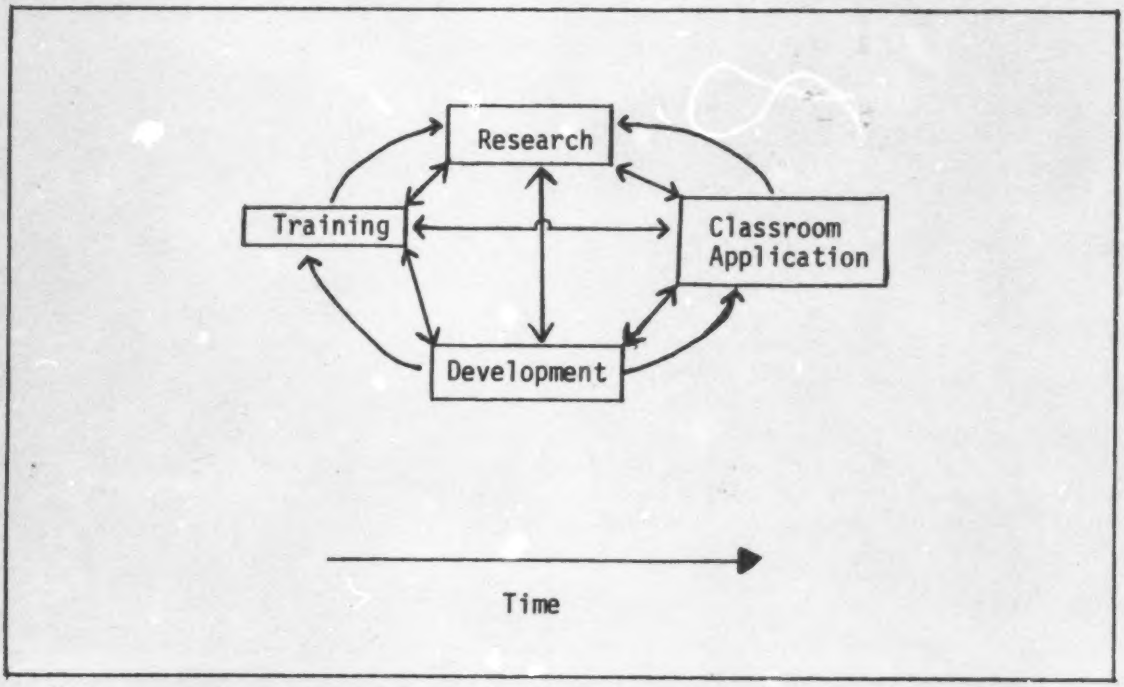
---

* Again, while we are not taking time to present the ideas here, Y. Bishop, S. Feinberg, and P. Holland present some statistical procedures that might be used to support this notion; see Discrete Multivariate Analysis: Theory and Practice. Cambridge: The MIT Press, 1975, also S. Feinberg, The Collection and Analysis of Ethnographic Data in Educational Research. A paper presented at the Workshop Exploring Quantitative/Qualitative Methodologies in Educational Research, Monterey, California, July, 1976; sponsored by Far West Laboratory for Educational Research and Development.

Basically, the model we are advocating consists of a team composed of teachers, researchers, and trainers. It is interactive in that all three constituencies participate equally in deciding what about teaching should be studied and how it should be studied. (See Figure 1). Having reached this decision, each member of the team undertakes tasks reflective of his/her own expertise and interests in order to answer the research question. Thus, teachers would inquire into their own teaching to determine the extent to which they and others already may be using those aspects of teaching on which the research is focused. Consequently, researchers would proceed to conduct the study, and trainers would monitor the data collection procedures used by teachers and researchers in order to establish ways of applying these same procedures in training. They can conduct small-scale pilot tests of some of these applications, and interaction with others on the team will enable them better to hypothesize possible outcomes of the research and their application to training.

FIGURE 1

Interactive Model of Research
Development, and Training in Teaching



From Ward and Tikunoff, An Interactive Model of Research and Development in Teaching, San Francisco: Far West Laboratory for Educational Research and Development, 1976.

As we have worked with developing the interactive model, four elements have made apparent the feasibility of its success. First, the questions to be asked grow from the concerns of the teachers on the team and are therefore immediately relevant and more useful. Past research and development efforts have seldom included this perspective, and thus the result often has been answers to questions unasked by teachers and therefore unusable by them.

Second, because the team operates on-site at a school, the concerns addressed are more likely to be relevant to the instructional priorities at that school than to national priorities set elsewhere. This is not necessarily to say that these priorities would differ, but that participation in setting priorities more often results in their ownership and in motivation to carry them out.

Third, all of the activities of the team occur concurrently, thus providing greater probability that findings and training procedures will be used more immediately. In the old, linear research and development model, the time lag was seldom less than 10 years between the inception of research, development of a product, and dissemination through trainers.

Fourth, the context-based observational procedures we have developed for use in our own work allow for the collection of more meaningful, qualitative data (Tikunoff, Berliner, and Rist, 1975; Ward, Tikunoff, and Lash, 1976; Tikunoff, Ward, and Lash, 1976). These include both participant observation by the teacher as (s)he teaches as well as non-participant observation by other(s) who watch him/her teach. Such a procedure allows for a unique inside-outside view of teaching without which the question of appropriateness cannot be pursued. Earlier in this paper we provided examples of such data from our work in discussing validity and reliability.

The application of an interactive model such as the one we have proposed attends to the concerns we have registered here. Given that teaching is viewed as a multimethod, multiperson, multisituation, multivariable activity, the collective efforts of a team of teachers, researchers, and trainers not only could attend to the assessing of teaching competencies but to the developing of training procedures to improve teaching. The collective expertise and concerns of team members can be focused on studying and improving teaching for a particular style. Within the framework of context-based teaching, this means collaborating to:

(1) identify the areas of teaching to be studied;

(2) selecting the indigenous and relevant teaching competencies, and building site-relevant matrices for these competencies;

(3) determining the degree of difficulty for performing that teaching competency given the multiplicity of context dimensions;

(4) establishing a procedure for assigning the level of performance for a teacher who operationalizes the given competency;

(5) multi-person observation of the teaching act to determine its appropriateness, and hence its validity;

(6) proceeding with data collection to establish inter-event and inter-teacher reliability for competencies that occur within similar context dimensions and that are the same and/or functionally equivalent teaching acts; and

(7) developing training procedures both to study and to improve those teaching competencies that are generalizable to the site(s) where the teaching will be carried out.

REFERENCES

Berliner, D.C. Impediments to the study of teacher effectiveness. Journal of Teacher Education, Spring, 1976.

Brophy, J.E. Reflections on research in elementary schools. Journal of Teacher Education, Spring, 1976.

Deutscher, I. Work and deeds: Social science and social policy. In W.J. Filstead (Ed.), Qualitative Methodology. Chicago: Markam, 1970.

Houston, W.R., et al. Resources for Performance-Based Education. Albany, NY: University of the State of New York, State Education Department, 1973.

McDonald, F.J. Report on phase II of the beginning teacher evaluation study. Journal of Teacher Education, Spring, 1976.

Merton, R.K. Social Theory and Social Structure(Rev.). Glencoe, IL: Free Press, 1957.

Patton, M.Q. Alternative Evaluation Research Paradigm. Grand Forks, ND: University of North Dakota Press, 1975.

Stallings, J.A. How instructional processes relate to child outcomes in a national study of Follow Through. Journal of Teacher Education, Spring, 1976.

Smith, L. Effective teaching: A qualitative inquiry in aesthetic education. A paper presented at the Workshop Exploring Qualitative/Quantitative Research Methodologies in Education. San Francisco: Far West Laboratory for Educational Research and Development, 1976.

Tikunoff, W.J., Berliner, D.C., and Rist, R.C. An Ethnographic Study of the Forty Classrooms of the Beginning Teacher Evaluation Study - Known Sample. San Francisco: Far West Laboratory for Educational Research and Development, 1975.

Tikunoff, W.J., Ward, B.A., and Lash, A. Proposed Research Design: A Case Study of Teacher Allocation and Utilization of Teaching/Learning Time During the Teaching of an Experimental Teaching Unit in Mathematics. Report D76-2. San Francisco: Far West Laboratory for Educational Research and Development, 1976.

Ward, B.A., and Tikunoff, W.J. An Interactive Model of Research and Development in Teaching. San Francisco: Far West Laboratory for Educational Research and Development, 1976.

Ward, B.A., and Tikunoff, W.J. The effective teacher education program: Application of selected research results and methodology to teaching. Journal of Teacher Education, Spring, 1976.

Ward, B.A., Tikunoff, W.J., and Lash, A. Proposed Research Design: A Case Study of Teacher Allocation and Utilization of Teaching/Learning Time in Mathematics During Regular Classroom Instruction. Report D76-1. San Francisco: Far West Laboratory for Educational Research and Development, 1976.

CONFERENCE SESSION

Beatrice Ward and William Tikunoff
Far West Laboratory For Educational Research and Development

Ward

I come to where I'm standing today, and to the thinking that is in the paper today, after eight years as a classroom teacher, four years trying to install innovative education programs that were individualized and about eight years trying to develop competency-based teacher...well, five years trying to develop competency-based teacher training materials and two and a half years in basic research on teaching. So that's my background.

Tikunoff

I represent fourteen years of classroom teaching from adult all the way up to kindergarten. I started with adults and worked my way up. I spent five years with developing competency-based teacher education. I guess I can rightfully claim some ownership of that whole movement and responsibility for one-half of the team that developed the model the Teacher Corps is now using in their programs. I have been involved at the Far West Laboratory for the past two and a half years in an attempt to bring teaching to research.

Ward

I'll start by simply presenting the dilemma that research on teaching is in in terms of the reliability issue and the validity issue, and Bill will talk a little about what we're doing as a result of that. When we say validity we mean are you measuring what you think you are measuring.

Reliability means consistency. It is a good word to use. It can be statistical reliability: test-retest; it can be interrater reliability: everybody sees the same at the same time. If you use the same observation form, and you have five people look at the same event at the same time--we've been pretty good at that. You can really do quite well with interrater reliability. We have been absolutely lousy on interteacher reliability. And we have been even more lousy on intrateacher reliability. Let me explain what I mean by those.

Interteacher reliability is three teachers teaching the same lesson to basically the same group of kids, and you're looking at the same kind of approach to teaching. Very low, really low.

Intrateacher reliability is teacher A on Monday, teacher A on Thursday, teacher A on Monday of the next week - even worse. I don't want to say any more than that. I just want to say the dilemma is that we can't do any of this very well at the point we're at.

Tikunoff

The reason is that research has failed to develop ways to show how to take an isolated competency and relate it to student outcomes. There are a couple of reasons for that. The first is when we spent some money as a federal government to catch up with the Russians. The same system for evaluating education was merely transferred from the system used to evaluate the development of missiles, and we bought that as educators. So we have ourselves to blame for the dilemma we find ourselves in. But even worse than that, those of us in research bought a similar model that the psychostatisticians adopted out of the world of natural science. People in educational psychology are beginning to say we have to begin to look at the way we have gone about doing things, because I have this strange feeling that psychostatisticians became arrested in this development of statistical models sometime around 1927. All research, therefore, has grown out of somebody's fascination with the way to analyze it. It is no wonder, then, that all the questions for which we have answers are questions that teachers aren't asking, and that is where we are now.

There are two solutions to this, we believe. One is to look elsewhere than within the world of educational researchers to get some answers. We at the lab have gone in two directions, one of which has been to move to applied statisticians outside the world of educational psychology, refreshing people like Steven Feinberg who is at the University of Minnesota. He tells me that anytime he has students who want to deal with quantitative counting, he insists that they deal with the qualitative aspects of the data that they are going to count. I'll give you a couple of examples, one of which is measuring the movement of lizards in the Bahamas. He took all of his students there for one month and they sat on the island day and night and watched the lizards and took field notes. The second example was the amount of time it took for a police officer to respond to a call in Detroit. He put all of his students in squad cars and they spent a full month in squad cars getting to know the data they were handling. These are the qualitative aspects.

About two and a half years ago we turned to the areas of anthropology and sociology to give us two techniques, one of which is called non-participant observation (somebody who is not involved in the act and sits and observes and takes notes). The other is participant observation--in this instance, the teacher. So we have turned to the qualitative aspects of data collection. That is not counting--that is sitting and trying to describe what on earth is going on, given an event. When we do that it is easy to find why, as we move from teacher to teacher, events look differently. The second major point is that we have gone beyond looking at one component to the next component to component analysis of the teaching act, primarily because we believe that one cannot look at the act of teaching and learning without looking at those context variables that affect what is going on. Therefore, even if you wanted to go from one teacher to the next, you're looking at totally different phenomena. We call this, to borrow Lou Smith's language, a multi-message, multi-person, multi-situation, multi-variable approach to what we refer to in the paper as a context-based teacher education.

Ward

During this last year, Bill and I spent six weeks in nine classes where we had non-participant observers and the teachers as participant observers. One of the things we were trying to do was look at the effect of kids' time-on-task. We were absolutely certain that high-achieving kids would spend more time on tasks than low-achievers, and we got fooled --they don't. It is that the high achievers accomplish a lot more during the time they are on the task. They do two or three things successfully at the same time.

Tikunoff

We had to look at the daily work of students in order to arrive at outcome data that made sense to us. Essentially what we had to do was assign a degree of difficulty to the problems being attempted and multiply that out by the number of problems the kids attempted and multiply that out by the number they actually accomplished. That seems to make more sense to us in terms of what kids are actually doing and what they are accomplishing. But to look merely at the engagement time of a high and a low youngster in the same setting, they look like they are spending an equal amount of time.

One of the other things we found, incidentally, is that the high kids in our target sample were apparently able to handle more than one stimulus at a time.

We've begun working on the notion that teachers ought to be researchers. I would hope those of you who are at teacher training institutions are training your teachers to be researchers, and by that I mean to approach the act of teaching with open inquiry and with a scientific mind. One of the notions we had was that we could take the methodology we use with ethnographers (people out of anthropology and sociology who observe) and use students to create protocols. We have been working with the University of Vermont to develop a system for training people to become observers. They are just now piloting that and are sending freshman students out into the schools to observe and then to develop these protocols. If they then can begin to ask questions about what they saw, they at least will begin to have more perspective on what occurs in the classroom.

Question: As a group, where do we go; and what do we do; and where should be put our priorities, given the next hour to work?

Ward

First, I think that as a training staff responsible for reliability and validity of our program, I'd probably get the reliability-of-instrument question out of the way and take care of that with a pretty traditional psychostatistical model. I'd probably do it using interobserver, or interrater reliabilities because you've got the highest chance of having them come out with a high amount of agreement that you can then use to support the instrument. I'd then go on with two things. I would want to be able to get as quickly as I could to some common agreement on the validity issue based on appropriate teaching. That issue directly leads you into the need to begin to build some type of context matrix. It

might not be as complicated as the one we've outlined, but you certainly need to get it more complicated than you now have it. I'd deal with that issue next. Having done that, I think then you'll get to a point where you can identify appropriate teaching with sufficient depth. Accuracy really isn't the word--it's depth. That is where I'd go with it. Don't you agree?

Tikunoff

I'll just add to that. I'd want more than one person looking at that performance. If a non-participant observer picked an act that looked inappropriate, the teacher often had very good reasons for doing what he or she did at that moment, and it probably was appropriate. That is not a reliability issue and it is not a validity issue. It is an interpersonal issue.

Question: You have been reviewing, as well as completing, a number of teaching effectiveness studies. Are there ideas you would like to share?

Ward

One of the things that has come out of this year's work is that whole issue of the functional equivalents, more than one way to do something may be equally productive for kids. We also are finding some things that maybe won't actually be what we want to hear in terms of what is productive for low-achieving kids, among them being the problems associated with the fact that in most classrooms a student is expected to attend to more than one thing at a time. It is very typical for a teacher to expect you to work in a workbook and listen too. The fact that students cannot do that is really a problem.

We have two measures that we are using this year in our research. One is the daily work outcome: how well the kid did with the daily work assignment he or she was given.

Question: How did we arrive at a daily work issue?

We're only half way there. We started by taking 264 different types of math problems which was what we had used in four days of observation in the nine classrooms. We had many more than that if you were to go within categories. We had a math specialist teacher and a math educator rank these on the level of difficulty of the problem for a kid to perform. They came out with thirty-four levels of difficulty which we forced them to collapse into twenty-four, and then they finally agreed they could collapse it to ten. The highest level of difficulty is a reading word problem that requires that you combine four or five math processes in order to arrive at an answer. The lowest level really is simple addition. This is fourth grade, and they weren't concerned about basic number concept.

Then, for each student's work sheet in the classroom for that day, each problem that was on it was scored for level of difficulty. In one classroom all thirty-eight kids had a different worksheet. In four or five of the classrooms, we had maybe six or seven worksheets that were done by different sets of kids. In two classrooms all the kids did the

same one. Each problem was scored. We scored whether the students tried to do it or not, and then if her tried to do it whether it was correct or incorrect. All of that is being put together to come out with a daily work score. The classes will be compared based on a level of difficulty score. We can talk about kids within classrooms now, but we want to be able to talk about classroom 1 compared with classroom 2 and we're now in the process of getting that scaled out so we can.

I think the process is generalizable. We were told by the people from California that they tried to get someone to do that with the achievement test items in terms of level of difficulty, and nobody would touch it.

Tikunoff

One of the other things that we are finding is this. It goes along with the functional equivalence notion. We are the mother and the father of something called experimental teaching units. Betty used them first in her dissertation, and then I developed four of them for the beginning teacher evaluation study for the State of California that Dave Berliner runs out at the lab. We've learned something there that I think is important for people. What these are, are curriculum packages that we can use to study the effects of teaching. This time when we did that, we insisted that the teachers teach all the objectives to all the children, and we got immediately into trouble. I think when we use those kinds of instruments we have to allow the differences to appear. The more constraints we place on teachers in situations that we try to control, the more we are going to affect the teaching and directly cause it to change in ways that we don't want it to change. We did two studies, fortunately, and we'll be comparing those data next year as part of our work. Where teachers were allowed to teach the way they wanted to, they did individualize, and they grouped, and they did what they would do naturally. Given the experimental teaching unit, not one of our teachers individualized initially, and all of them started out using whole group instruction. I just pass that advice on to you who are attempting to do, as best you can, what they do normally, so that you can watch them as they work.

Ward

The California State Study we are working on started out with the grand idea that you could go out and correlate all these marvelous big variables and ended up where it is now, in fifteen to twenty classrooms doing very detailed studies working with maybe five kids in the classroom, watching the teaching as those kids are involved with it. I would say that is basically where we have come from. The grand correlation of whether the teacher uses praise or not was highly correlated but meaningless because you didn't even know whether the teacher should have used praise, and one study would have a high correlation and another a low. The odds are now, if you had the data to go back into those studies, you'd find it was appropriate in one class and inappropriate in another. So that's really where we are.

We also have the bias going that random samples of teachers are not what you want to look at or work with. You want to work with the most effective classrooms you can find. Commonalities across effective classrooms can be far more meaningful than variations among the effective

and ineffective.

Tikunoff

That is another important point. In looking for a sample of effective teachers, we spent about $195,000 for the BTES Study, identifying forty classrooms, twenty of which were more effective, and twenty of which were brand X. This last time we spent about 195,000 hours (not money) negotiating with one school district. We went to the group of teachers who served as their research board, and we said: "Would you help us with a research study. What we're looking for are ten of the most effective teachers that you have in teaching mathematics at the fourth grade level." They went out and did their homework and they came back and said: "listen, we're not comfortable. There is one person here we've got to check out. Can you work with nine?" And we said: "We can work with nine." So, we worked with the nine teachers. On the basis of what we've done we want to go back now and doublecheck to see whether or not in fact they look like effective teachers when you compare them with everybody else in the district. I think sometimes we fall into the trap of thinking that we need quantitative evidence that we are in fact working with something called a "good" teacher. I have a hunch that any one of you here can walk into a classroom and tell me if there is good teaching going on in there or not.

Ward

One of the other things that has really opened up our thinking is to move out of working with education statisticians. Working with Steve Feinberg who is just a statistician has really opened us up. I used to be worried if I had an empty cell; Steve considers that the most important data source you can find. He also has methods and the books that are referenced that I can't get through, frankly, because my stat background is not large enough. They allow you to get into fifteen-to-twenty- dimension studies if you want to; and you can, in fact, do that with one classroom, depending upon how you identify what you variables are.

Tikunoff

We are also beyond the statistics. We have been to the sociologist, the psychologist, the anthropologist. We're fascinated with Etology, the study of how organisms adapt to their environment, primarily human organisms. Educators have not, as have the natural sciences, developed a rigorous, systematic, way of going about describing what it is we are about. The reason people in statistics can point to a formula and have any idea that it is going to turn out okay in the end is the fact that they have several hundred years of existing descriptive data. What we did was borrow the statistics sometime about the turn of the century (because there were a lot of people who were screaming for all their kids to come to school) in order to sort out which kids ought to go on to higher school; and we invented the achievement test. Nobody has stopped to look, historically, at why we are where we are and who we are. I think it is about time that we began to spend our efforts in collecting descriptive evidence of what schooling is like. We are beginning to do that with the work that we are doing.

Ward

In many ways, what we are talking about now sounds like the old 1930's applied research that we went through. I think the difference between what is being talked about now and what was going on during that period is that we in fact do have some analyses and quantitative techniques that can actually apply in the classroom.

SOME RANDOM REMARKS ABOUT WAYS AND MEANS OF INSURING RELIABILITY AND VALIDITY IN COMPETENCY ASSESSMENT

Patricia M. Kay

Bernard M. Baruch College
The City University of New York

One of the unique advantages of being a reactor rather than a presenter at a conference is that it is rarely necessary for a reactor to prepare an extensive paper prior to presentation. Reactors also have a great deal of flexibility in the manner in which they choose to fulfill their function. One the one hand, (a) they may be highly critical of the work being responded to - in effect, picking apart research methodology, demolishing theoretical arguments, pointing out unwarranted assumptions, and deriding the feasibility of practical suggestions. On the other hand, (b) they may be highly laudatory - praising and expanding upon the highlights of the presentation. Sometimes, of course, reactors do neither. In that case, (c) reactors may take the opportunity afforded them by program design to set forth their own totally independent set of ideas or point of view. And sometimes they do (d) all of the above or (e) none of the above. Any of these choices may be appropriate responses. The problem surfaces when the paper, not required before the fact, is required after. That problem is reflected in the first three words of the title of this paper.

At the Plattsburgh Conference I attempted to do several things, which I shall try to refer to here. When the conference was being designed, this paper was intended to be a response to the presentation by Ward and Tikunoff entitled "Specifying Ways and Means of Insuring Reliability and Validity in Competency Assessment." However, early on in the conference itself, data about the concerns of the participants vis a vis reliability and validity of assessment systems for teacher education programs was collected. In part, my "reaction" to the Ward-Tikunoff presentation was to deal with some of those concerns and to attempt to point participants towards appropriate resources to deal with them. Although the questions raised by the conference participants may have been unique to that group, it seems likely that anyone reading this far in this volume may be looking for similar assistance and have some of the same concerns.

The concerns of the conference participants about the reliability and validity of assessment systems for competency based teacher education programs fall, roughly, into three categories that might be labeled:

1. "Who will believe us?"

2. "I don't have enough technical knowledge about reliability and validity at my fingertips to feel adequate to the task."

3. "I don't know about you folks, but at my institution there aren't enough people, money and time to tackle all this stuff. HELP!"

First, let's consider the "Who will believe us?" or the program approval category of concerns. In New York State, the development of CBTE programs has been mandated and is being monitored by the State Education Department's Division of Teacher Education and Certification. Conference participants were eager to know what the Division would consider appropriate evidence that particular assessment systems were adequately reliable and valid. What kind of reliability and validity evidence would SED be looking for; how much would be enough; would data need to be gathered on each instrument in an assessment system?

On the surface, these concerns seem to be related primarily to a worry over program approval that is specific to New York State. In point of fact, however, these are the same worries that beset any group of educators undertaking program renewal or redevelopment and who wish to gather evidence about the effectiveness of their undertaking. Unfortunately, there is no simple or single answer to that problem. The nature and extent of the validity and reliability evidence needed to make a particular program's assessment system "believable" by others depends, to a large extent on the nature of the competencies being assessed, the form of the assessment, the context in which the assessment takes place, and the use to which the assessment data will be put.

Yarger (1975) attempted to sort out some of the complexity inherent in the question of the quality of evidence that is needed in different situations. He addressed the question of how good is good evidence by suggesting that the goodness or "hardness" of evidence may be rated on a scale going from MUSH (meaning not so hot) to ROCK (meaning pretty good). He then pointed out that the same bit of data could be thought of as very hard evidence to one set of persons (administrators needing to make a decision), and a very soft evidence to another set of persons (researchers needing to confirm a theory). His obvious conclusion is that one person's ROCK may be another's MUSH.

William Boyd, who represented the Division of Teacher Education and Certification at the conference, addressed this problem area most appropriately when he stated that the division would be looking primarily at the kinds of efforts that colleges were making towards establishing the reliability and validity of their instruments and that SED's interest would be more in the processes undertaken than in the final statistical results. The response is particularly appropriate in view of the fact that many institutions are forging ahead with truly unique approaches to assessment systems. There are no fully developed, highly reliable, completely validated systems to be used as models. Different approaches to competency assessment may demand different approaches to methods of validity and

reliability data collection and even some different notions about what reliability and validity mean. The approach presented by Ward and Tikunoff at the conference is a perfect example of how reliability and validity must be thought of as an integral part of the total approach to competency assessment development.

The Ward-Tikunoff paper presented a number of intriguing ideas that seem to have potential appeal for teacher education faculty. Their basic thesis emphasizes that the world of teaching is a multivariate world and that any attempt to assess teaching competency must use a multivariate approach. Teacher educators have probably always been aware of this most salient fact. Teacher behavior researchers have been trying for some time and through a variety of means to identify and define those many variables and the relationships between and among them. (If we knew with any good degree of certainty what they were, the validity problems would not be quite so serious.) It is only recently, though, that the real demands of this assumption have been addressed in both research and training practices.

The situation is taken a step further by the important contribution made by Ward and Tikunoff. Their approach to building competency assessment systems is both multivariate and practically implementable for teacher education programs. They start with the notion that teaching acts may be appropriate in some contexts and not in others (something that teacher effectiveness research is finally beginning to show.) The assessment trick, then, becomes one of recording the contextual variables as well as the teacher acts involved and obtaining judgments on the appropriateness of those acts in those contexts. In this view, appropriateness equals validity,and reliability is defined as the consistency with which teachers perform appropriate acts.

These researchers have, in effect, given teacher educators a model for building assessment systems that seems to be highly consistent with previous inclinations of teacher trainers and which lends an air of legitimacy to the use of expanded ratings and judgmental data that have typically been the assessment tools of teacher evaluation. It is likely that many teacher educators would opt for this manner of developing assessment instruments. There are, however, cautions that should be exercised which need to be pointed out.

The basic problem is more related to the "old fashioned" (but still good) notion of validity than to the legitimacy of using this particular approach. At this point in time, there are serious limitations to the state of knowledge about what kinds of teaching behaviors or acts have predictable effects on the learning of pupils in various contexts. Since this is the case, there is probably little harm in using assessment systems that depend heavily on judgmental data. About the best that can be done is to rely on the accumulated wisdom of the profession. Ultimately, however, the appropriateness of certain acts in specified contexts depends on whether or not children learn what they are supposed to learn as a result of those acts. In that case, whether or not certain teaching acts are appropriate becomes an empirical question involving "harder" data than consensually validated, multiple judgments. Seven hundred modern day

judges and the ghost of Columbus could agree that the Nina, the Pinta and the Santa Maria landed in America in 1776, but that wouldn't make it true. We have better evidence than that.

Admittedly, the example about the discovery of America is on the absurd side. It does, though, serve to illustrate that we need to keep in mind that common sense and experience or authority based judgments are not always borne out by data. There is a tendency, particularly in those social sciences dealing with everyday phenomena about the behavior of people, to assume that consensus equals truth. Frequently, in the fields of psychology and sociology, that which has been hypothesized through casual observation and "common sense" has been confirmed by data ... but not always.

This limitation should not be taken as a warning against using the "context based" procedures suggested by Ward and Tikunoff. Neither should it be inferred that teacher educators should do the research necessary to accumulate the evidence for ultimately defining truth about teacher effectiveness (although Ward and Tikunoff suggest some very exciting possibilities about how teachers and teacher educators may become involved in that process.) Rather, these comments should serve as a caution that using the suggested procedures could result in a kind of static institutionalization or standardization if the persons doing the judging do not keep themselves knowledgeable about research results as they accrue. Any approach to competency assessment that is heavily grounded in judgmental data demands that those making the judgments keep abreast of developing knowledge in the field. The more highly inferential the judgments, the greater the need to know what the research tells us.

The second category of concerns expressed by the conference participants at Plattsburgh was related to a continuing need for technical assistance. Questions in this category were numerous and the following are illustrative. "How does one go about establishing the validity of the kinds of instruments that may be used ... especially with reference to attitude measurement?" "Is content validity appropriate and, if so, how do we establish that?" "We need to know more about indices and methods of calculation ... especially for instruments that are to be used in classroom settings." What are the techniques for establishing the validity of rating scales?" "We need guidelines for the development of observation systems." Once again, these concerns are not unique to the group of special education teacher preparation consortia members who were attending the Plattsburgh conference. These same concerns arise across the country wherever teacher education faculty are struggling with the problems of competency assessment.

Most teacher educators are not specialists in the methodologies of either teacher effectiveness research or educational measurement and statistics - to say nothing about the highly specialized combination of those areas that seems to be demanded by problems of evaluating reliability and validity of competency assessment systems. This is probably the one area where continuing assistance for teacher education program development groups may have a high payoff for very little investment.

There are probably two alternatives available. The first may not be feasible, and the second may be onorous, but both have the potential for improving assessment systems by quantum leaps. Alternative one is to find a consultant who will be willing to work with a program development group as the competency assessment instruments take shape and begin to acquire a history of use. Such a person should be well grounded in the appropriate methodologies, but should also be committed to taking a bird's eye view of the development of particular systems and suggesting appropriate strategies for dealing with assessment related issues as they arise. The investment here would be an outlay of cash, but the advantages that would probably accrue are that a great deal of very valuable faculty time could be preserved for working with students and similarly valid enterprises.

The consultant, in this case, need not be external to the institution, but may be. On practically every campus housing teacher education programs, there are people who are specialists in measurement and research methodologies. Sometimes they are unable or unwilling to lend their expertise and sometimes (often) there are no monies available to support even a very limited degree of external consultation. In this case, then, teacher education program development groups must undertake a "do-it-yourself" project. For those in this latter category, there are many resources that are both valuable and available. The following references are strongly recommended for them.

1. Gage, N. L. Handbook of Research on Teaching. This volume has been, for many years, the "Bible" for those involved in the many issues attending research on teaching. There are a number of chapters that contain very pertinent information for those who would increase their knowledge about the problems of instruments in competency assessment. Of particular interest to teacher educators who must deal with field based assessment systems is the chapter by Medley and Mitzel about the development and use of observation instruments. Although it was written several years ago, it is still regarded as THE important reference on the topic.

2. Travers, R. Second Handbook of Research on Teaching deals with many of the same issues as the "first" Handbook. It adds to the contents those developments of the intervening years and is probably one of the most complete, authoritative, and up-to-date works that is currently available.

The two handbooks, along with a basic graduate level text on measurement and evaluation in education and psychology, which is likely to be found on most teacher educators' reference shelves, should serve to get the process started. For those who are a bit further along, perusing reports of current research studies for methodological implications as well as research results may prove fruitful. That is, after all, the jumping off point for Ward and Tikunoff in describing the approach they suggest to competency assessment development. The methods they describe for assessment in CBTE programs were not developed for that specific purpose ... they grew out of their own research. Other large scale research studies may contain similar implications. The largest compendium of such studies that is available at the present time is reported in the Spring, 1976 volume of the Journal of Teacher Education.

And, finally, those looking for assistance in this area should consult the proceedings of several conferences that have been devoted specifically to this topic. In addition to the portions of these proceedings that deal with problems of validity and reliability, particular reference is made to the Houston (1974) and Dixon (1975) volumes dealing with research and evaluation of teaching competence.

The third set of questions raised by the conference participants all seemed to have the same message: "Wow! This is an incredibly complex task. We don't have the answers and neither do all the experts that have been brought in to assist us. Even if we did know exactly what we were doing, it is an undertaking that demands money, time, resources, facilities and people that we don't have at our disposal. Yet, the process of examining our goals has been rewarding, the problems of improving instructional contexts of teachers has been exciting. Now, what can we do?"

The very first thing that people who are in this state need to realize is that they are not alone ... and that the task will only overwhelm them if they allow it to. Almost all of the programs that have made significant progress toward becoming competency based have done so without extra funding or people and under serious time constraints. The secret seems to lie in getting organized, developing a plan of action that all involved can agree on, and then slowly and methodically attacking the problems as they arise. No one who admits to the eventual payoff of this kind of teacher education approach thinks that it can be accomplished overnight ... and, luckily, no one is demanding that that occur! "Traditional" teacher education programs evolved over a period of many, many years. Surely, the ability to cope with the issues raised by competency based programming and assessment will develop and mature with time and attention. Meanwhile, those who find some merit in the ultimate goals may find excellent examples of current "coping strategies" in the work reported by Shearron that is going on in Georgia, by Baird about the program developments at Brigham Young University and by their many colleagues who so willingly share their ideas at conferences such as the one held in Plattsburgh.

REFERENCES

Gage, N. L. (Ed.) Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Houston, R. W. (Ed.) Competency Assessment, Research, and Evaluation: A Report of a National Conference. Syracuse: National Dissemination Center for Performance Based Education, 150 Marshall Street, Syracuse University, Syracuse, N.Y. 13210, 1974.

Travers, R. (Ed.) Second Handbook of Research on Teaching. Chicago: Rand McNally, 1973.

Yarger, S. J. From rocks through melon to mush: The place of the teaching center in research and evaluation. In G.E. Dixon (Ed.) Educational Comment: Research and Evaluation in Operational Competency-Based Teacher Education Programs. Toledo: College of Education, University of Toledo, 1975.

DESIGNING A SYSTEM FOR MANAGING ASSESSMENT IN COMPETENCY BASED TEACHER EDUCATION PROGRAMS
(A Resource Paper)

J. Hugh Baird
Brigham Young University

INTRODUCTION

A story is often told of the airplane pilot who interrupted his passengers to announce that he had good news and bad. The good news was that they were in a jet stream and were making excellent time, the bad news---that they were lost. Today, many teacher education programs are lost - they are not moving their students toward an identified goal. Competency-based teacher education is an attempt to prepare prospective teachers in terms of specified observable goals. It is a significantly different kind of program. Stan Elam has defined it as follows: a program in which "performance goals are specified, and agreed to, in rigorous detail in advance of instruction. The student must either be able to demonstrate his ability to promote desirable learning or exhibit behaviors known to promote it. He is held accountable, not for passing grades, but for attaining a given level of competency in performing the essential tasks of teaching."[1]

Five elements are now generally agreed to be essential to competency-based programs.

1. Teaching competencies to be demonstrated are role-derived, specified in behavioral terms, and made public.

2. Assessment criteria are competency-based, specify mastery levels, and made public.

3. Assessment requires performance as prime evidence; takes student knowledge into account.

4. Student's progress rate depends on demonstrated competency and is,therefore, self paced.

5. Instructional program facilitates development and evaluation of specific competencies.

In the writer's experience there are four other critical elements.

1. They must contain worthwhile content.

---

[1]"Performance-Based Teacher Education - What is the state of the art?", Stanley Elam, A.A.C.T.E. 1971

2. They must be humane.

3. Eventually competency-based teacher education programs must be based on more than assumptions about training experiences and teacher skills and outcomes in children and therefore -

4. The training system must be self-corrective. It must be able to seek and respond to feedback from the users of the system. This quality of being self-corrective may not be required in competency systems but because we are going to make mistakes as we identify competencies to be attained by trainees, because the profession does not yet have all the data needed on which to base and establish competencies, we, therefore, need to build a self-corrective system which constantly gets feedback from the community, from children and parents, from the users of our product, and modify our system in terms of that feedback.

Assessment will be defined here as the generation and use of information that supports decisions relative to the trainee or the program. Most programs in operation began with, and still spend most of their energy attending to, assessment problems regarding trainee performance. However, because of the absolute necessity for all present day CBTE programs to be self-corrective, these materials are organized so that questions about trainee and program assessment are considered simultaneously.

The following eight questions need to be answered specifically and in some detail, as an assessment management system is planned. Ideally, decisions regarding these questions will be made by those persons to be affected by the decisions. Your time today should be spent in thinking about the questions and writing tentative recommendations of answers which will result in the best assessment system for your program.

1. What information (data) is to be used in making decisions about trainee and/or program?

2. Who uses these data - for what purpose?

3. Who is to collect the data?

4. How often are data to be collected?

5. What instruments and procedures are to be used in collecting data?

6. When are data needed (how often)?

7. In what form are data to be kept and/or displayed for those who use it?

8. How are decisions and judgments communicated to those persons who need to know?

It will soon be evident to serious planners that each of the above questions is interrelated, but more important, each is a general question implying many specific questions.

As you struggle with questions regarding assessment there are two principles which I value, and which you may choose to use as guides.

Principle 1: Parsimony - In making a decision, don't collect any more data, or involve any more persons than are absolutely necessary.

Principle 2: Involvement - Those affected by a decision should be involved in making it.

You will notice that each of these principles tends to serve as a check on the other.

WORKTEXT

Although all questions are interrelated, question one is probably the place to begin. Some space is provided after each question for your tentative notes, ideas, and recommendations.

I. What information is to be used in making decisions

A. About Trainees

1. Learner Outcomes

a. What do they need to know?

b. What teaching skills must they possess?

c. What pupil outcomes must they achieve? (under what conditions)

d. What other qualities must they possess? What other requirements must they meet for certification?

2. Learner Characteristics

a. Background

b. Physical Characteristics

c. Scholastic ability

d. Personality

e. Attitudinal

3. Descriptions of the setting in which teaching has occurred.

   a. school

   b. classroom

   c. pupils

B. About Program

1. Staff

   a. training

   b. performance

   c. etc.

2. Instruction

   a. modules

   b. scheduling

   c. sequence

   d. etc.

3. Information transmission

4. Logistic support

5. Record keeping

6. Assessment and research

7. Counseling

8. Decision making

NOTE: Data - about program element might include descriptive and evaluative data - how it works, how effective it is, what it costs in time and dollars.

II. Who used these data - for what purposes?

A. What information about his performance does the trainee need and when does he need it?

1. To help him improve

2. For other reasons

B. What information do trainee's supervisors need and when?

1. To help trainee improve

2. To help supervisor improve

C. What information do other instructors need?

D. What information do those who make judgments about trainee's competency need? When? (Some programs designate a panel of faculty - not to include the trainee's supervisor - who makes final judgment on a trainee's competency.)

E. What information do program administrators need to make judgments about the program?

III. Who will collect data? (Will this be teachers and supervisors mainly, or will "third party" observers also be used?)

IV. How often will data be collected? (Sampling, validity, cost are important here.)

V. What instruments and procedures are to be or can be used to standardize, simplify and accurately collect data?

VI. How often are data needed to be displayed for those who use them? What data are to be available?

A. On demand?

B. Once per week - for staff meeting for instance?

C. At the end of each semester?

D. For yearly evaluation and revision sessions?

VII. In what form are data to be kept and/or displayed for users?

RESOURCE MATERIALS

(Editorial note: The following materials were provided to illustrate the application of the principles and processes presented in Dr. Baird's resource paper and in the related conference workshop session.)

4.0 Competencies are treated as tentative predictors of professional effectiveness, and are subjected to continual validation procedures.

Indicators:

4.1 Program includes research component to validate competencies.

Personnel and resources are specifically assigned to this component.	Planned validation effort is evident.	Some validation testing is done.	No planned validation effort is evident.

4.2 Competency statements are continually analyzed and revised.

Statements are systematically reviewed for possible deletion or revision for each training cycle.	Some competencies are revised or deleted for each training cycle.	Competencies are treated as permanent objectives.

10.0 Learner progress is determined by demonstrated competence.

10.1 The student is knowledgeable of the general nature of competencies and criteria used to determine the extent to which performance approaches professional standards for acceptability.

Student describes competencies and the standards for acceptability.	Student not able to describe competencies required or criteria that are acceptable.

10.2 Learner progress records are adequately detailed in terms of the competencies to be acquired.

Learner progress records are adequately detailed, in terms of the competencies to be acquired.	Learner progress records not kept on file.

10.3 Learner progress records are used to chart future programs' directions.

Learner progress records are frequently used to chart program direction.	Learner progress records are seldom if ever used to chart program direction.

10.4 The demonstration of progress in acquiring the competency is the focus of attention in determining the extent to which the learner is experiencing success.

Success is determined by extent of progress in acquiring the competency.	Success is determined by some other criterion such as amount of knowledge acquired, or number of activities completed.

10.5 The instruction management system makes provisions for students to be working at various points of development concurrently.

Instruction is modularized and organized to be carried out individually or in small groups by variable scheduling techniques.	Instruction is based on the assumption that all students should acquire the same learnings at the same time.

11.0 The extent of learner's progress in demonstrating competencies is made known to him throughout the program.

Indicators:

11.1 Learner progress records are maintained and available to all concerned (learner, instructors, counselors).

Learner progress records are accessible, adequately detailed, and open to himself, instructors, and counselors.	Learner progress records are inaccessible, inadequate, and/or closed to students.

11.2 The instructional staff (instructors and counselors) and learner periodically review progress records in conference.

Student progress conferences are held frequently.	Student progress conferences are nonexistent.

11.3 The instructional management system provides for the frequent and/or continuous updating of the student's progress records.

Progress records updated on a continuing basis.	Progress records if available are only updated at infrequent (i.e., semester end) periods.

11.4 The student is provided with opportunities to acquire skill in analyzing and evaluating his own professional behavior.

In addition to being provided with information about his progress, the student is helped to acquire skill in analyzing his own professional behavior.	Little or no attention is given to the analysis of the student's progress, and none in helping the student acquire this skill himself.

12.0 Instructional specifications are reviewed and revised based on feedback data.

Indicators:

12.1 Specifications for the instructional system are explicit and all concerned (students, instructors, counselors, instructional professional services personnel, etc.) are aware of these specifications.

A list of specifications for the instructional system is published.	Neither specifications nor policies concerning the instructional system have been recorded, much less made known to those involved.

12.2 Procedures have been established for having students assess the instructional system.

On a frequent, periodic or continuing basis, students are asked to react to the effectiveness of the procedures used in the instructional system.	No attempt is made to obtain students' reactions to the instructional procedures.

12.3 A wide range of data is considered in the analysis of the instructional system. (Student time, instructor time, instructional resources, management needs, learner performance, etc.)

An extensive collection of data is used for the analysis of the instructional system.	No attempt is made to analyze the operation of the instructional system.

12.4 Data obtained from the analysis of the instructional system as provided by student feedback are used to revise the system.

On a frequent, periodic or continuing basis, the instructional system is revised from data provided by student feedback.	No systematic or regular attempt is made to revise the instructional system. Changes are made primarily on demand from some condition or authority.

Assessment

13.0 Competency measures are related validly to competency statements.

Indicators:

13.1 A listing of performance indicators is included with each competency statement.

Multiple indicators are present for all competency statements.	Few competency statements have multiple indicators.	No competency statements have more than one indicator.

13.2 Indicators are logically related to competency statements.

all competencies	some of them	none

13.3 Measuring instruments are logically related to indicators.

all competencies	some of them	none

14.0 Competency measures are specific, realistic, and sensitive to nuance.

Indicators:

14.1 Competency measures discriminate between learners who demonstrate and those who do not demonstrate competency.

All measures most of the time	Most measures most of the time	Undetermined

14.2 Measures assess consistency of performance over time.

Always	Usually	Seldom

14.3 Reliability of instruments is known and high.

Computed for all instruments and high	Some instruments	Not known

14.4 Procedures for measuring competency demonstration are specified so as to assume quality and consistency.

Generally followed and known by data collectors	Procedures not specified, known or followed

14.5 Data collection procedures require realistic time and resource expenditures by students and staff.

Realistic	Unrealistic

16.0 Data provided by competency measures are manageable and useful in decision making.

Indicators:

16.1 Data are collected and stored in an easily retrievable form.

Data on competency measures are collected and centrally stored	Some data are collected, storage not planned or centrally located	Not collected or not stored

16.2 Data are reported at pre-specified decision points.

Reports are helpful to decision makers	Some reports are made as a result of special needs	No reports are generated

16.3 Data are used in making programmatic decisions.

Data are generated as a basis for decision making	Occasionally data are used as a basis for decision making	Not used

16.4 Data collection and analysis procedures are feasible in terms of time, personnel, and resources.

Efficiently handled within resources	Collected but seldom used because procedures are cumbersome	Burden is on program, or not collected

16.5 Data are easy to interpret.

Format of data analysis is clear	Not easily interpretable

17.0 Competency measures and standards are specified and made public prior to instruction.

Indicators:

17.1 Competency measures and standards are in a written form.

For all competency measures and standards	Some are available	None are written

17.2 Competency measures and standards are specified in advance.

Yes	For some	NO

17.3 Students can describe competency measures and standards.

All known to them	Some known	Unknown to students

17.4 Procedures for demonstrating competencies are known to students and faculty.

Known to all	Known to some	Unknown to students

Governance and Management

18.0 Policy statements are written to govern, in broad outline, the intended structure, content, operation and resource base of the program.

Indicators:

18.3 Policy decisions are supported by and made after consideration of data on program effectiveness and resources required.

Data are collected and systematically stored and considered in reviewing, changing or creating policies	No research base exists for policy decisions Policies are the result of power relationships and personal opinions

22.0 Research and dissemination activities are an integral part of the total instructional system.

Indicators:

22.1 A research strategy for validating and revising the program is operational.

Written procedures, hypotheses, data; systematically applied	Some efforts to study results of program	Not being done

22.2 Reports of completed studies are used in revising program.

Numerous written reports available, used	Data or unwritten reports available	No reports

22.3 Research management system is operational.

Yes, comprehensive, workable, working	Some processes, not systematic	Not operational

22.4 Procedures for sharing results with other programs and for obtaining their reports are operational.

Regularly shared with at least two programs, some sharing with ten others	Haphazard sharing of results with other programs	No relationship other than casual ones

22.5 Staff can describe the research strategy, on-going studies, and conclusions of previous efforts.

All staff	Some	Only for studies he is engaged in

Indicators:

24.1 The program was planned as a totally integrated system.

Total program designed prior to independent parts	Courses compiled into a program	Independent parts grouped together and called a program

24.2 The program is operated as a system.

Decisions reflect consideration of the total system	Many isolated independent decisions

24.3 Management is by objectives.

Yes	Somewhat	No

24.4 Evaluation system provides continual feedback to assess objectives; achievement for various sub-systems.

Data available and used; program revised	Data occasionally used	None operational

24.5 When making decisions on one phase of the program, impact on other sub-systems is calculated and considered.

Always	Sometimes	Never

24.6 The sub-systems are continually being modified.

Yes	Somewhat	No

24.7 Harmony in principles among various sub-systems is apparent.

Internal consistency easily apparent	Consistency can be generally identified	No consistency, or not considered

24.8 The program is continually evaluated against the actual professional needs, and refined based on feedback.

Formal review structure operational; changes continually being considered	Program not amenable to modification

ISTEP Supervisor_____
Academic Supervisor_____ BYU Number_____
Cooperating Teacher_____ Soc. Sec. #_____
Public School_____ Major_____
Subjects Taught_____ Grade Level_____ Married_____ Single_____
_____ Grade Level_____ Class: Jr.___ Sr.___ Gr.___
_____ Grade Level_____

Pass-Fail
476 P F Number of Hours ____
479 P F M
H. Ed. 362 A B C + -

Grades
301 A B C + - 403 A B C + -
377 A B C + - 415 P F
310 A B C + - 479 P F M
H. Ed. 362 A B C + -

Pass-Fail
476 P F Number of Hours ____
479 P F M
H. Ed. 362 A B C + -

Grades
301 A B C + - 403 A B C + -
377 A B C + - 415 P F
310 A B C + - 479 P F M
H. Ed. 362 A B C + -

Student Signature_____ Supervisor Signature_____ Date___
Supervisor Signature_____ Grades Submitted_____ Date___

NOTE: If you are registered for 479 with your major department, circle M.

*A-100 _____	*D-220 _____	G-254 _____	*G-470 _____
*A-110 _____	*D-221 _____	*G-255 _____	G-473 _____
*A-120 _____	*D-222 _____	G-256 _____	G-474 _____
*A-130 _____	*D-223 _____	G-257 _____	G-475 _____
*A-140 _____	*D-224 _____	G-258 _____	G-476 _____
*A-150 _____	*D-225 _____	*G-259 _____	*G-477 _____
*A-160 _____	*D-226 _____	*G-261 _____	G-550 _____
	*D-327 _____	*G-262 _____	
B-205 _____	D-427 _____	or	*H-280 _____
B-206 _____	D-428 _____	*G-263 _____	*H-380 _____
B-303 _____	D-520 _____	G-264 _____	*H-381 _____
B-305 _____		*G-265 _____	H-382 _____
*B-307 _____	E-231 _____	G-266 _____	*H-383 _____
B-405 _____	*E-332 _____	G-267 _____	*H-384 _____
*B-408 _____	E-333 _____	G-340 _____	H-480 _____
*B-410 _____	E-530 _____	*G-350 _____	H-481 _____
B-500 _____		G-351 _____	*H-483 _____
	F-240 _____	G-357 _____	H-580 _____
*C-211 _____	*F-241 _____	G-367 _____	
*C-212 _____	*F-242 _____	G-368 _____	J-200 _____
*C-213 _____	F-343 _____	*G-369 _____	J-201 _____
C-214 _____	*F-344 _____	*G-370 _____	*J-202 _____
C-315 _____	*F-445 _____	G-371 _____	J-203 _____
C-414 _____	*F-446 _____	G-372 _____	J-204 _____
C-416 _____	F-540 _____	*G-375 _____	J-205 _____
C-511 _____		*G-385 _____	*J-300 _____
	*G-250 _____	*G-386 _____	*J-301 _____
	*G-252 _____	G-466 _____	*J-400 _____
			J-590 _____

Student Teaching: Beginning Date Projected_____ Actual_____ Ending Date Projected_____ Actual_____

KEY TO READING CONTRACT:
*Required for B or pass/fail. Required for an A--starred objectives, additional major ques', project, outstanding student teaching, complete within 2 semesters.
D-427 _____ Initial Contract D-427 _____ Dropped through renegotiation
D-427 _____ Added through renegotiation

EDUCATION-BIOLOGY 276

(2 copies: Instructor, Student)

Name _____

Pages	Assignment	Attempts	Date Achieved
5	10 Camp Out	_ _ _ _ _	_____
7-10	20 Keeping Up-To-Date Reading	_ _ _ _ _	_____
11	30 Operating Instructional Equipment	_ _ _ _ _	_____
13	31 Preparing Instructional Materials	_ _ _ _ _	_____
15	32 Bulletin Boards	_ _ _ _ _	_____
17-18	33 Field Trips	_ _ _ _ _	_____
19-26	34 Collections	_ _ _ _ _	_____
27-28	35 Laboratories	_ _ _ _ _	_____
29-37	36 Teaching Games	_ _ _ _ _	_____
39	37 Innovative Science Programs	_ _ _ _ _	_____
41-42	38 Value Clarifying Activities	_ _ _ _ _	_____
43-48	39 Individual Differences	_ _ _ _ _	_____
49-77	40 Classroom Management	_ _ _ _ _	_____
79-111	41 Writing Instructional Objectives	_ _ _ _ _	_____
113-137	42 Performing A Task Analysis	_ _ _ _ _	_____
139	43 Evaluation Theory	_ _ _ _ _	_____
141-159	44 Concept Learning Theory	_ _ _ _ _	_____
161-171	45 Interaction Analysis	_ _ _ _ _	_____
173-200	46 Using Questions in Teaching	_ _ _ _ _	_____
201-216	50 Inductive Micro-teaching	_ _ _ _ _	_____
217-250	51 Memorization Micro-teaching	_ _ _ _ _	_____
251-259	52 Changing Affect Toward Living Specimens	_ _ _ _ _	_____
261-267	53 Lesson Teaching	_ _ _ _ _	_____
269	60 Concept of Teaching	_ _ _ _ _	_____

We are attempting to evaluate the syllabus for this class. Listed below are each of the sections of this syllabus. In the first column, please indicate if you studied that section or not. If you studied that section, indicate if you felt it was adequate or not, and if you felt that it was not, please indicate what you think could improve that section in the space provided, such as eliminate, condense, re-write, clarify, expand, etc.

	Studied?		Adequate?		Recommended Changes
	Yes	No	Yes	No	
Table of Contents					
Course Goals					
Assignment Sheet (Eval. Theory) p. 1					
To Determine Type of Test pp. 3-4					
Types of Measuring Instruments p. 5					
Norm and Criterion Referenced p. 6					
Evaluation pp. 7-10					
Practice Items pp. 13-21					
Formative-Summative Evaluation p. 23					
Attribute Sheet p. 25					
Instruction/Examples pp. 27-35					
Practice Items pp. 37-42					
Test Items pp. 43-47					
Attribute Checklist p. 49					
Essay Test Items p. 51					
Oral Test Items p. 52					
Developing Questionnaire pp. 53-57					
Observational Testing Methods pp. 59-60					
Norm Setting Evaluation Sheet p. 61					

	Studied? Yes	Studied? No	Adequate? Yes	Adequate? No	Recommended Changes
"Effective Communication" Sheet p. 61					
Evaluation and Mastery Learning pp. 63-69					
Evaluative Purposes p. 71					
Designing and Using Evaluation Materials p. 72					
Assignment Sheet "Preparing a Test" p. 73					
Assembling the Test pp. 75-76					
Diagnosing and Prescribing p. 77					
Examples of Diagnostic-Prescriptive Examinations pp. 79-105					
Assignment Sheet "Interpreting Student Test Performance" p. 107					
Norm and Criterion Referenced Grading pp. 109-111					
Assignment Sheet Affective Evaluation p. 113					
The Place of Affective Learning pp. 115-119					
Non-Cognitive Outcomes pp. 121-122					
Course Related Behaviors p. 123					
Examples of Affective Instruments pp. 124-144					
Assignment Sheet Multiple Talent Evaluation p. 145					
Be Talent Developers pp. 147-150					

UNIT REPORT FORM

Unit 10

Instructions: Please report on the learning activities you engaged in as you studied this unit.

	Very Helpful-Worthless					Minutes Spent	
1. Reading Text	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more
2. Lecture on X	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more
3. Lecture on Y	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more
4. Lecture on Z	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more
5. Film "_____"	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more
6. Help Session	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more
7. Supplementary Reading A	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more
8. Supplementary Reading B	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more
9. Conference with Instructor	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more
10. Studying with other students	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more
11. Other:_____ _____	5	4	3	2	1	☐ 0-30 ☐ 60-120	☐ 30-60 ☐ 120-more

Signed_____

PROJECT 4 - CONFUSION QUESTIONNAIRE

OBJECTIVE

The common college experience is altogether unlike the ISTEP experience. The demand of self discipline on the part of the student in individualized instruction is a substantial change from the spoon-fed approach.

There is ample reason to believe, if you listen to departmental secretaries, that many entering ISTEP students are confused about various aspects of the program.

It is the objective of a properly structured questionnaire to determine where and how serious the confusion is.

Areas where confusion commonly surfaces are:

1. How to work on modules
2. Grading procedures
3. Overall requirements of program
4. Expectations of student performance on service projects
5. Expected performance during student teaching
6. Requirements for completing a quest

It is apparent that confusion about some of the above items, and others would not arise until later in the term. Questionnaires are thus given during the third, seventh, and eleventh weeks.

* * *

INSTRUCTIONS TO STUDENTS

ISTEP is currently undergoing an evaluation which will isolate its strengths and weaknesses as a teacher training program. Appropriate changes will be made, based on data collected.

This questionnaire is part of the data collection process and is designed to help the researchers know <u>what</u> it is about ISTEP that confuses you, <u>when</u> that confusion occurs, and <u>how long</u> it lasts.

On each of the items listed in the questionnaire, indicate:

A. <u>Whether you have felt confusion</u> about the item by using the following scale:

1 = no confusion
2 = some confusion, but not enough to affect my work
3 = enough confusion to slow down my progress
4 = extreme confusion - I can't do anything until it is cleared up

B. When that confusion occurred by marking the timeline from the first week in the semester you felt the confusion to the week when your confusion on the particular item disappeared.

C. What confused you in the area (it is important that you write specifically what it was that confused you).

D. Whether or not you attended a seminar or help session to help you with this objective. If you DID attend a seminar or help session, circle the number of the question.

Example: 1. Requirements for Quests

a. 3 (amount of confusion)

b. week #

1 2 3 4 5 6 7 8 9 10 11

c. How do I write a Quest Proposal?

QUESTIONNAIRE #1

Students responded to the following questions by circling a number1-11 on a scale opposite each question. Example: Question 1

1. Registration Procedures

a. _____

b. 1 2 3 4 5 6 7 8 9 10 11

c.

2. Where to go for help
3. What is an objective
4. How to filfill objectives
5. Requirements for Quests
6. Keeping a Journal
7. How to find staff, materials, returned objectives, etc. in the LRC and McKay Building
8. How to use the syllabus
9. Contracting and renegotiating
10. Goal Setting
11. Seminars
12. Campout
13. Requirements for Service Project
14. School Visits
15. Concept of Teaching
16. Personal Development Unit
17. Adolescent Development Unit
18. Health Unit
19. Administrative Aspects of Teaching
20. Curriculum Unit
21. Instruction Unit
22. Classroom Management Unit
23. Unit flow charts

QUESTIONNAIRE #2

Students responded to the following questions by circling a number 1-11 on a scale opposite each question. Example: Question 1

1. How do I use a form 1 and what is all the information on it for?

a. _____

b. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |

c.

---

2. How to find staff, materials, returned objectives, etc. in the LRC-McKay Bldg.
3. How to prepare for the simulation game.
4. How to use Administrative Aspects of Teaching Evaluation sheets.
5. Position and situation assignment in the Administrative Aspects of Teaching Simulation Game.
6. Newspaper editorials in Simulation Game.
7. Ideas to be Learned in Simulation Games. (Administrative Aspects)
8. Procedures and rules in the Simulation Game.
9. Required and optional objectives list.
10. Having a vision of what ISTEP is and should be.
11. Introduction Unit
12. Personal Development Unit
13. Adolescence Unit
14. Health Unit
15. Curriculum Unit
16. Instruction Unit
17. Classroom Management Unit
18. What do you do when you can't find material in the LRC?
19. Choosing and using learning activities.

QUESTIONNAIRE #3

Students responded to the following questions by circling a number 1-11 on a scale opposite each question. Example: Question 1

1. What is required to pass Micro-teaching lesson?

a. _____

b. 1 2 3 4 5 6 7 8 9 10 11

c.

---

2. Concept of Teaching
3. What happens when you finish this semester and what happens if you don't?
4. What is required on the Individualized Unit, Objective G-477?
5. What do you do when you can't find material in the LRC?
6. Inductive Teaching
7. Expanded Inquiry, Objective G-476
8. Introduction to Syllabus Unit
9. Personal Development Unit
10. Adolescent Development Unit
11. Health Unit
12. Curriculum Unit
13. Classroom Management Unit
14. How to write a programmed unit
15. Problem-Solving Theory
16. Classifying Questions
17. Analyzing Objectives
18. Evaluation Theory
19. Choosing and Using Learning Activities
20. Reading the flow chart in Unit G Instruction

CONFERENCE SESSION

J. Hugh Baird
Brigham Young University

Let me tell you a couple of things about me so that I can help you most. I have been working in a competency-based program. It is not special education. I have little or no expertise with regard to special education concerns directly. But with regard to evaluation, with regard to program design, I have had some experience.

In 1966, three of us went to our department chairman and told him we were very dissatisfied with the way we were talking about individualizing instruction for groups of seventy or eighty students, and asked for permission to start a program for prospective secondary teachers which would be more consistent with what we were teaching. He said, "Excellent! Go at it. No, we don't have any money to give you to develop this program. Well, no; as a matter of fact, you can't have load credit. You don't have any students, and so there is no way to give you load credit or released time. Well, I really don't know what kinds of changes could be made in University policy or whatever. You'll just have to test those as you go. But you have my support." And so we began.

On the way back from Chicago one winter we identified into a tape recorder about sixty-five things that we thought our graduates, beginning secondary school teachers, ought to be able to do. The next semester we got eight students, all social studies majors that semester, who were willing to experiment with us. The three of us and eight students and sixty-five objectives went to work. No materials. No program. And the program developed. . . has become an experimental program, a good, rich source of resources, ideas, and people in which to try almost anything we want to try. The University has been extremely supportive. The approach has always been, "Well, why not? How can we arrange it so that you may do that?" And without that we couldn't have developed a program, because there have been zero federal funds. Over a ten-year period we have had about $11,000 in state funds, and most of that went into buying video recording equipment and video tapes. The University in that eleven-year period has given us about the equivalent of two full-time faculty members for one year. We have had an increased budget of about $14,000 spread over ten years to buy a few supplies and materials. Now, those were our resources. The State Department has been the same way. They allowed those eight students to be certified on our word that they were competent. They have allowed and encouraged our students ever since.

We have been able, as a result of our experiences, to help other institutions develop their assessment programs or their management systems.

I remind you again. It is not in the context of special education that I have been working. It is with a great range of secondary education students majoring in music and math and art and history and physical

education and English and home economics and business education and whatever.

The thing I'd like to do for the next few minutes is to distribute a paper. It has two to three pages of ideas about assessment systems. And then the rest of the paper is kind of a work text. A lot of questions are asked with space for you to write answers. And I'd like to give you about a minute and a half to read these first two or three pages, kind of with me, and then walk you through that work text part, giving illustrations as we go. When you get to the top of page four, stop.

Let me help you now by just pointing out a couple of things that I would hope to stress in this first two or three pages. The notion that training systems must be self-corrective doesn't simplify your task. Most programs get started looking at the evaluation of students: and then, almost as an afterthought, if there is money and time, we worry about evaluating our program, following-up on our graduates, or checking out to see how well we are doing. We need to be validating our program. We need to be checking up on our graduates to find out what they are doing. So I take a strong position that when you start you assume that your program is going to be self-corrective, very responsive to data that is being collected all along the way. It is in that context that I talk. Now that complicates your life, but it will eventually bring you out of the tunnel at the other end, I think. It was in that context that I was willing to start with eight students and no program. Some people and the theorists would probably tell you that you have to wait two or three years. You have to do all the conceptualizing, and you have to simulate your program; and you have to do this and this and this before you dare put a student in it. I say you can't really know about a program <u>until</u> you get students in it. Probably the greatest plus for you people in New York is that you are running on-going programs. You have the trainees there to work with, and so you begin: get your programs going; and then collect data on how effective they are, and what the influence of what you do is on the outcome, and modify it. A perfectionist just goes batty in that kind of a situation, and you may, too; but I don't know of any other way that we are going to learn than to build self-corrective elements into the programs, and then just go with them. I am not worried at all about that. I believe that wherever we are, we can do at least as well as we have been doing over the last ten or fifteen years. And I have every hope that we can do better with the energies we are spending.

These forms we are developing (the evaluation procedures, and instruments) help us generate data which hopefully will be used in these decisions. There is one principle that is not stated here. It is one that I learned in Oregon, and it goes like this. Data are collected to be used, but data do not make the decisions. <u>People</u> make the decisions, based on the best they know - hopefully based on good data, but <u>you</u> make decisions. Hopefully you increasingly clean up your data and help it become more powerful. To get a form, an evaluation sheet, and to arbitrarily say, "You have to earn a score of 50 to 78 on this evaluation sheet before you are judged to be competent," is not right, in my judgment. You collect the data, and then a <u>person</u> makes a judgment. Is that adequate or is it not? Do you see the difference, as opposed to saying, "Well, Bill, the form says you got 68 and you need 70, and so you are not competent." - thus letting the form make the decision? Now

that may be a fine point, but somewhere in your program you are going to have decisions to make where you will decide the data are not trustworthy, or you blew it in collecting. . .you got the wrong kind of data, or you do not have enough, or something else. And you still have to make the decisions. Hopefully, you will make them using the best data; but the data does not make the decisions for you.

Also, you ought to be involving your students in making some of the decisions that you will be asked about here. You ought to certainly be involving all of your faculty. Both the faculty out in the public schools and the faculty on the campuses. They ought to be involved in a lot of these decisions. They are going to be affected directly by them. For you to sit here in the Ivory Tower of Plattsburgh and make those decisions independently of them is not right. That will get you in trouble, and so that principle ought to be applied wherever you can.

There are two principles at the top of page four. The researchers sort of cough and sputter and turn green with one of these two principles, the Law of Parsimony. The researchers sometimes are inclined to say, "Collect all the data you can. You may want to use it someday." That kills you off. That will just do you in. Involve as many people as you possibly can, and then make this as lean a program as you can make it. Apply this Law of Parsimony and collect what you need, of course, and think through that very carefully, but don't collect any more information than you need.

For instance, we used to ask the student a whole lot of demographic information about himself: social security number, major and minor, who the supervising teacher was, what school he was assigned to. Every time they turned in a route slip, students had to fill that dumb stuff out. We were using that, and it was just getting in the way. So this form has become very lean and it asks for only the information that we need to process. You just be as stingy as you can with students' time and your time, and with your energies. Keep your program lean. Make it do what it needs to do, but nothing extra.

The second principle concerns decisions related to what information you need in order to make decisions about trainees. In other words, what information do we need about what our students can do? Can they ask appropriate questions? Can they operate equipment? Can they write lesson plans? Can they do these kinds of things? In our program, we collect data about what our students can achieve as they work with children in the schools. Our program is one which requires students to get results with public school children before we will recommend them for certification. They have to make a difference, a positive difference in the schools. It is not enough to do all the things teachers do - those things have to get results with kids before we will certify our students. And so, what do we need to know about our students' ability to get results with children?

And then the last concern: What information do we need about our students' affect, about their attitude toward themselves? Other questions are also important. What background information do you need on your students: about their physical characteristics, their scholastic abilities, their attitudes? We need to know about subsequent employment. Students

at Brigham Young University, over half of them, come from outside the state of Utah. They come from New York and Dallas and Denver and Portland and St. Paul, and all the little towns of the nation too, and they go back to their communities to teach. Wouldn't it be very helpful if we knew, for instance, that over 50 percent of our students were finding jobs, after training, in small, rural towns, working with high Indian populations? Wouldn't it be very important that we know that kind of information as we design our training program? That would result in a different kind of program than if our graduates were all ending up in inner-city schools. The program is going to be different, and so it may be that you need some information about subsequent employment.

You may want to collect data that describes the setting in which your students do their student teaching. The context in which they work will be very important to you as you make program improvement or curriculum changes. What information about the program is needed in order for you to make management decisions? What information about the staff do you need to collect? For instance, are all your students being supervised by doctoral candidates? Are the college supervisors people who have had rich public school background, rich special education background in the schools; or are they theoreticians who have never set foot in a special education classroom? You might want to know that about your staff as you consider staff evaluations of students, or staff recommendations for modifying the program.

Another thing about staff is worthy of mention. I remember seeing a chart coming out of an elementary program a couple of years ago, and it showed how differing staff members used a particular evaluation form. There was one staff member who consistently rated his students at the top of the form. He gave them all 5s on a 5.0 scale, on everything on the form. Student after student, school after school, year after year. There were other faculty who never gave a higher score than a 3 on that form. Do you see what I am saying? You need to know about how your staff is behaving. You might need to know that some of your staff members have a reputation of being excellent when it comes to helping students do certain kinds of things, and they are seen by students as being inhibitors or the problem itself in other areas of their training. If I were directing a program, I would want to know those kinds of things about staff so that I could make appropriate decisions based on valid data in designing a program.

Okay, let's keep moving. Other elements of the program you may need to collect data to make wise decisions about include instruction. How powerful are your modules, your lectures, or your whatever? Do your learning activities do what you designed them to do? Does this particular text really teach what you want it to teach? How about scheduling? Are you making staff members accessible? Are they around when the students are around? Interestingly enough, our staff members have to change their schedules mid-way through every semester. As the bulk of their students go out into the public schools and leave the campus, their schedules have to change. They have to have office hours late afternoons or they are just not available. There are many other quesions listed in the paper, and you may even want to add questions of your own.

COMPETENCY ASSESSMENT: A PERSPECTIVE
(Based on transcription of the conference session)

Robert Brinkerhoff
Evaluation Research Center
University of Virginia

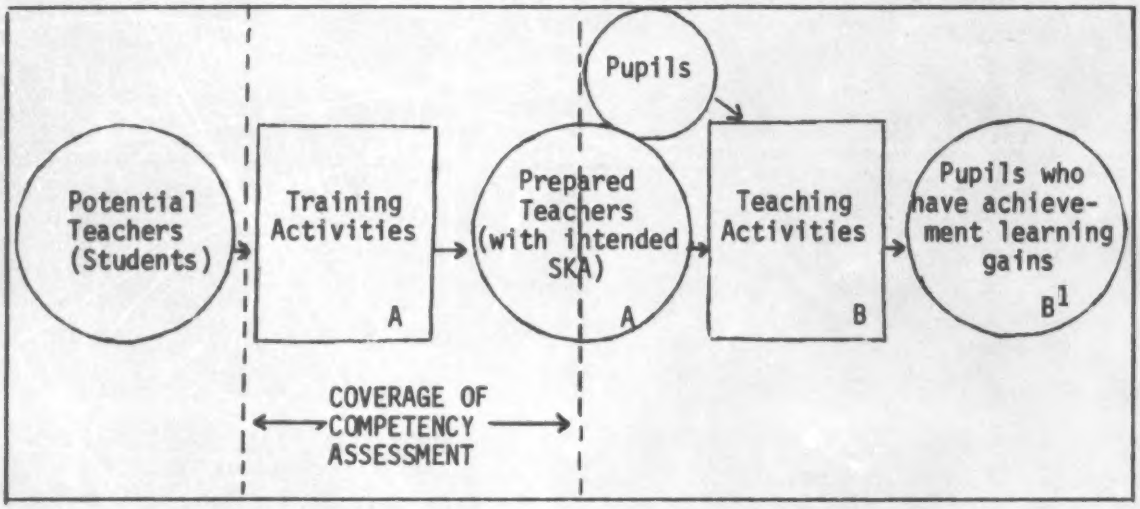
The need to assess competencies in a teacher training program is not unique to the CBTE (or PBTE) based program - it is a real and pressing need of any training program that intends to engage in evaluation. In a CBTE program, the training paradigm itself assumes evaluation (assessment of pre and post competency levels.) A "regular" (non-CBTE) program makes no implicit evaluation promises. Yet, when evaluation enters into the regular program formula, then no difference exists between the two kinds of program - each needs to determine, in as objective a way as possible, just what it is that its trainees have learned: what they know, how they feel, or what they can do.

Limiting the Assessment Problems

Competency assessment is not an end in itself. Like many other activities, it has several uses. When we talk about assessing competencies (when we talk about anything, for that matter), it is useful to know just what it is we mean by it. It will behoove us to proceed to define more precisely what it is that we expect a competency assessment to do for us. Competency assessment is, no doubt, a problem for many teacher educators; it may help to remember that competency assessment as an activity is supposed to be a solution to a problem - the problem of evaluating our teacher training programs.

The diagram in Figure 1 shows a very simplified analysis of the rationale and function of a teacher preparation program.

Figure 1
Rationale and Functions of a Teacher Preparation Program



As shown in Figure 1, training programs (A above) exist to convert students into teachers: these teachers then instruct pupils in schools (B) where the pupils (hopefully) achieve some learning objectives. There are several key assumptions here, only some of which are within the scope of competency assessment as an evaluation method. These assumptions are:

1. The skills, knowledge, or attitudes that are the objectives of the program (A^1) are valid and worth achieving, that is they are capable of bringing about intended pupil changes (B^1).

2. The students who enter the program (A) do not already possess the skills, knowledge or attitudes (SKA) that represent the objectives of the program (A^1).

3. The training activities within the program are potent; that is, they can bring about the changes necessary to convert the student to a teacher.

4. The teacher who exits the program has in fact, achieved the SKA that are the objectives of the program (A^1).

5. The teachers who have left the program in fact use and apply (at point B) the SKA they acquired (as A^1) in the program.

Whereas assumptions number 1 and 5 are critically important, it is this writer's contention that their testing, or proof, is not the intended aim of competency assessment as applied in program evaluation. It is, of course, true that competency assessment procedures and results could be useful in testing either assumption (first #5, then #1). Nontheless, the testing of these assumptions is not the responsibility of program evaluation, or thereby of competency assessment in the context of program evaluation. Assumptions 2-4 are considered the proper focus of competency assessment. From attention to these areas, the use of such assessment procedures can determine the status of incoming students relative to the program objectives, and monitor the achievement of these objectives throughout and at the end of the program. Student achievement data thus collected can serve two major functions: (1) individually, it can be used to make decisions affecting the learning program for that student; and (2) as group data, it can be used to determine the effectiveness of program activities and elements.

At this point, it is probably necessary to provide a definition of competency assessment. In general it is a process; a process that produces information (data) about the extent to which students have mastered particular skills, knowledge or attitudes. Thus, competency assessment is a procedure engaged in to answer such questions as:

- How well did students learn what we expected them to learn?
- How did they progress through the program? What did they achieve?
- What in the program worked? What didn't?
- What can students do when they leave the program?

There are several dimensions of this definition worth further investigation. First, I have defined assessment as a process. Thus, to begin to do it, is to do it. This means of course, also, that it is not a product, nor an end point; one doesn't really ever finish it, either. It is a way of getting somewhere, and not a place to get to.

Another point worth mentioning is that assessment is a method for making decisions - it is a way of basing decisions, or judgments, on some data about reality. Important here is to note that assessment is not intended to replace judgment, or decision making; rather, it is a way of getting some information with which to make a decision, a judgment.

Essentially, one can view the assessment process as a question-answering method, or as a way of helping to fix belief relative to a question or concern. For example, in a teacher training program, we might pose such evaluation questions as:

1. "How well can student A design a lesson plan?"

2. "To what extent can graduating students diagnose learning needs?"

3. "How effective is the practicum program in teaching questioning procedures?"

Assessment, or evaluation, demands that these questions be answered in a particular way. They are answered not by conjecture or reflection, but by deriving judgments from a body of data. Thus, in order for the evaluator (or assessor) to answer these questions, he must obtain some data about reality, or performance. This data is not a direct reflection of reality, but some codified abstraction of it, collected through some sort of measurement process. This entire process is pictured in Figure 2.

Figure 2

Assessment as a Question-Answering Method

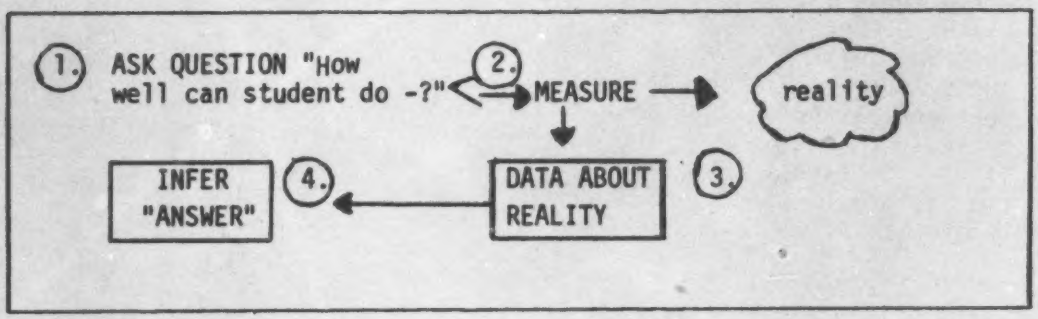


Figure 2 is intended to represent the major steps in doing competency assessment, and to demonstrate the fact that "answering" the competency assessment question is both inferential, and somewhat indirect. The data that derive from measurement will never represent the totality of the competency concept being gauged - i.e. one inevitably works from a sample of information abstracted from a reality, and from less than the total amount of information which could be seen to represent the concept. Let's look at an example.

> Example: If I am interested in the extent to which a student can perform a learning diagnosis, I might identify as a subcompetency, the ability to write an evaluation summary. I might determine that a reasonable measure of this competency could be obtained from an expert analysis of an evaluation summary report actually produced by a student (a work sample). Of course, I cannot expect to judge absolutely every possible variable or dimension of that summary report, but will settle on relatively few key aspects of it; define them as accurately as possible; then design a sort of rating instrument that will quantify the judgments rendered relative to these key aspects. Obviously, I've gone through a number of analytic steps and reductions to be able to get at some data - and clearly, I don't have anywhere near all the data I could possibly get to answer my original competency question. I end up looking at (judging) a very small "piece" of the total reality I set out after; from this small sample, I hope to make a valid inference about the existence, or extent, of some larger quantity.

To the practicing competency assessor, the principles demonstrated in this example contain both good news and bad news. The good news is: you can't possibly measure everything about all the competencies in your program, and you don't even need to try. The bad news is: you've got to measure some, and since we're involved in inferential analysis, there's alot of room for error. Before moving on to discussing the major assessment steps and some ways of reducing the error probabilities, there are two more points to make relative to the overall assessment process.

First, data don't answer questions; people do. All the data in the world about a competency are, alone, insufficient to determine that competency's presence, or value. Some person will use that data to make a judgment about the competency. Competency assessment techniques are not intended to replace human judgment in our programs; they are intended to make more systematic and objective the judging process. The data collected by measurement are used, or appealed to, as a "belief-fixer". The answer to the competency assessment question comes in the form of "I (or we) believe the competency is demonstrated to X extent, because these are the data we have."

The second point deals with the many analytic steps one must go through in order to arrive at a set of measurable performance indices. The assessment procedure begins with deductive analysis; one "tears down" a larger order concept, or competency area, in order to arrive at, finally, a set of performance indices sufficient from which, when observed, to infer the amount or value of the original concept. The deductive analytic process is necessarily reductionistic - one ends up with more specificity, but perhaps less meaning. If one is to remain

aware of the larger order of things, and of the place of the smaller things relative to it, these analytic steps must be carefully pursued, and faithfully documented.

This whole process is a bit like trying to determine the overall contents of a large box by plunging one's hand in and grabbing some small bit, looking at it, then guessing at what, and how much of what, that box contains. The trick is to first determine what you'll feel around for in that box, where, and how you'll know when you've got hold of it. Then, when you've got that bit out and are looking at it, you will want to know just how it might relate to the larger content - what bit you've likely got, how it relates to the whole, what portion or aspect it represents, and so on. If you've done that thinking beforehand, you're much more likely to come up with the bit you were feeling for, and you'll be able to make a better guess as to the whole contents.

Some Practical Steps for the Competency Assessor.

This next section will pose some general steps that one can pursue in the competency assessment struggle. Each is first listed, then individually explained in some more detail. Remember: these aren't steps for determining what competencies should be taught, nor are they steps for doing competency-based teacher education (Please refer back to the section on "assumptions" if those statements surprise you.) They are steps that a person can take in order to begin to collect some data about the extent to which students have, or have not, achieved the intended set of learning objectives of some instruction.

The general steps are:

Step 1. Identify the general competency (or ability area) that you wish to evaluate (normally one of critical importance to your program, or to the success of those who leave the program).

Step 2. Analyze (define) the general competency into a set of 3-5 dimensions - the major defining characteristics of the general competency.

Step 3. Determine a "measurement mode" for each dimension - i.e. a way of collecting data for each dimension.

Step 4. For each dimension, define a set of performance indicators (i.e. "measureables") that are susceptible to measurement by the mode selected in step (3).

Step 5. Develop (or select) an instrument that will record and categorize the performance indicators defined in step (4).

Step 6. Determine a standard or criterion for the measured performance.

There follow after these steps, of course, the processes of data collection, analysis and reporting, and finally the organization of a usage structure which ensures that competency assessment information can be effectively used in the program.

Each step is further defined and discussed below.

Step 1. Identify the general competency to be evaluated.

Given that you can't - and probably don't need or want to, anyway - measure every learning outcome of your program, a specific focus has to be identified. Many programs may be just starting into the competency assessment "game", and it is probably sensible not to bite off too much to begin with. The development of a system to collect performance information is best done incrementally; as time goes by and experience deepens, it will expand to cover more competencies, and to measure them more accurately. Most programs, too, have already identified lists of intended student learning outcomes (competencies). These lists, along with other information about the design of the teacher - training program should be analyzed to determine which competencies are most critical; if you could only measure one kind of student ability, what should it be? Once (at least) of these should be selected to begin with.

The competency selected should be both critical to the program and sufficiently general that it represents as hefty a "piece" of the overall set of intended learning outcomes as possible. For example, something as general as "...can adequately diagnose pupil learning needs", or "...can establish rapport with pupils" is about right. A statement such as "...knows how to teach is probably too general, and "...can name three behavior modification strategies" is too specific.

Step 2. Analyze the general competency into a set of three to five major defining characteristics.

Once a general competency areas has been selected, the next task is to analyze it into several more specific dimensions. The purpose of this step is to arrive at a more specific definition of the competency - to identify those abilities or whatever, that constitute the more general competency. Basically this step is needed to define what we mean by the competency.

For instance, if we had begun with the original competency of "...can adequately diagnose pupils learning needs", we might, completing Step 2, decide that this competency consists of:

a. is knowledgeable of general diagnostic tests
b. can select proper diagnostic measures
c. can administer ____ diagnostic technique
d. can interpret diagnostic information

From this example, you can see that the more general competency is defined as a set of four (4) dimensions. Notice that each dimension is

more specific than the beginning competency, yet that it relates to it. Notice, too, that not each of the dimensions is necessarily a "behavior" (such as [a]); no need to worry about this - the point here is simply to define. The set of dimensions should, as already mentioned, be more specific than the original competency; they must also be comprehensive. I need to ask - is this all? - if a student could do each of these things (dimensions), would I be satisfied that the student had the competency? If my answer is "no", then I need to ask "what else?", adding that other dimension to my list.

Step 3. Select a measurement mode for each dimension.

The purpose of this step is twofold. First, it will, obviously, give us a measurement mode for each dimension. It will in addition, however, greatly economize the next step; it will focus the further analysis of the dimensions, and save alot of analytic work. Any measurement task will require analysis of general concepts into measurable, or observable, performance indicators*, or variables. This can be long and expensive work, and usually nowhere near all of these performance indicators end up being measured, anyway. A big reason they don't (or won't) is that the way we set about to measure will allow us to observe some of them, but not others. The point here is that, in measurement, the variables eventually measured are the result of an interaction of the analyzed meaning of the concept and the measurement method being used.

A practical technique for arriving at - selecting - a measurement mode is to list, on the vertical column of a matrix, possible ways of collecting data (as has been done in Figure 3). These may not be all possible ways of measuring things, but might represent ways which your program is currently collecting data, or methods for data collection that are reasonable and affordable for your program.

Figure 3

Measurement Mode List (partial)

- observation of students in practicum (class/on-job)
- survey of site supervisor's perceptions
- self-report
- analysis of work-samples
- interviews
- objective tests
- observation of structured tasks
- etc.

*Robert Mager has discussed ways of doing this, as have others.

The next part of this step involves adding, as column heads across the top of this matrix, the dimensions of competencies which you've decided on. One then completes the matrix by checking (in the appropriate cell) which dimensions might be amenable to data collection by which measurement modes.

In general, measures of competencies should be as specific and as behavioral as possible. Data about competencies are most useful to a training program when they represent quantified behaviors rather than quantified judgments or opinions. The measure should allow one to reconstruct the behavior; when it does not, feedback is relatively nonutilitarian - it can tell a student that someone (a rater) didn't like what he did, but it won't tell him what he actually did, or did not do. The implications for training are obvious.

It is likely, of course, that perhaps most any mode might yield some kind of data pertinent to any given competency; there are certain criteria* that can and should be applied:

1. Directness - The mode should yield information as directly related as possible to the competency in question. That is, we need to limit the number of inferential steps to be traversed between the information produced by the measurement (data) and the competency being measured. Below are three examples of measurement modes of varying directness all related to the competency "Ability to construct a lesson plan":

   - In-direct - Check the learning of pupils exposed to receiving a lesson based on the lesson plan produced by the student.
   - Less In-direct - Observe a lesson being delivered by the student.
   - More Direct - Complete a content analysis of a lesson plan produced by the student.

2. Objectivity - The measurement mode selected should require as little subjective judgment as possible. This criterion is fairly self-explanatory, and will relate to both the mode selected and the measurement instrument devised for that mode.

3. Cost - All measurement costs time and money, and also may entail social or psychological costs. In general, the cost factors need to be balanced against the foregoing two criteria, and against the potential costs of not having objective and direct data on which to base a decision.

---

* The problems of reliability and validity are of course critical; but have not been dealt with here. The criteria discussed are, in fact, related to technical and theoretical concerns of measurement; these criteria are intended to be practical considerations, and when met, will improve both validity and reliability.

Step 4. Analyze variables that are performance indicators.

Now that a competency has been analyzed into its major dimensions (defined), and at least one measurement mode identified, one needs to determine the specific variables or performance indices to be counted, rated or categorized. These variables need to be as specific as possible, and uni-dimensional; from them, hopefully, one can write items for an instrument. The variables - or indices - are those specific behaviors or bits of information that comprise an operational definition of a dimension of a competency. Here is an example:

Competency: Ability to plan a group lesson
Dimension: Ability to construct an adequate lesson plan
Variables: (i.e. some of those things that constitute an adequate lesson plan)

- inclusion of an introduction
- inclusion of behavioral objectives
- inclusion of an assessment technique
- content analyst's rating of adequacy of above three segments
- references to prior lesson
- reference to literature, a theoretical basis
- procedure for grouping of pupils
- etc.

Step 5. Construct or select an instrument.

This is a crucial step in the overall process, and of course cannot be anywhere nearly fully dealt with here. Books have been written on the subject, and for some the reader is referred to any of the major psychometric texts. There are, however, a few practical considerations of note. Instrument development is, like the whole process of competency assessment, an incremental activity. One cannot progress until one begins - my point throughout has been that it really doesn't matter as much where one does begin; a start, no matter how primitive or unsophisticated, is a necessary precondition to improvement - and certainly better than no start at all.

To begin with, an instrument is a relatively simple device; it is a means of recording some observations or perceptions, usually with predetermined categories. If one has completed the previous five steps, one is very close now to writing an instrument - some way of recording data on variables determined in the previous step. For purposes of this discussion, it is useful to think of any instrument as having three major parts. These are:

1. directions to the respondent or recorder
2. items
3. response choices, or formats

Specific criteria for parts (2) and (3) will vary with the type of instrument being used. There are, however, general criteria of some practicality. First, directions should be concise and clear. They should tell

the respondent why he's being asked to do something, exactly how to go about completing the instrument, and what to do with it when he's done. In many cases, it should inform him as to what will happen to data he produces, and often should include a mention of protection of privacy rights. Finally, the language of the directions should be appropriate to the respondent.

The items written should be at least one item per variable, and preferably more, for reliability's sake. The items should be unidimensional (i.e. contain only one variable) as shown in the example below.

Uni-dimensionality:

Bad: "does the plan contain specific and behavioral objectives? ☐ yes ☐ no

Good: "does the plan contain written objectives? ☐ yes ☐ no

If yes:
. are the objectives stated behaviorally? ☐ yes ☐ no
. are the objectives specific? ☐ yes ☐ no

The items must also be appropriate to the response choice. Further, they must be clearly written and unambiguous. They also must define any terms that one might reasonably consider confusing to the respondent. Remember: the item has to mean the same thing to each respondent.

Response choices should be consistent with the item, and internally consistent - that is, they must also be uni-dimensional. For example:

Bad: "Do you like prunes?"
☐ yes
☐ sometimes
☐ never
☐ no

Good: "How often do you like prunes?"
☐ never
☐ sometimes
☐ always

or Good: "Do you like prunes?"
☐ yes
☐ no
☐ can't say

Responses should also be inclusive, but mutually exclusive. That is, the respondent must be provided with enough response choice to cover any possible response, but he must be clear as to which response choice is appropriate (as above, if I sometimes like prunes, do I check "yes", or "sometimes", or even both?).

In general, the instrument should be easily read, understood and completed. The item and response choices must be appropriate to the respondent, and place as few demands on him as possible. The easier the instrument is to complete, the more likely it will yield reliable and

valid data. Finally, it should require as little inference as possible on the part of the respondent.

Step 6. Constructing a standard or criterion.

The instrument constructed in the previous step represents, finally, an operational definition of some aspect of the competency begun with. Thus, the body of data collected by the instrument stands at the far end of a long chain of analysis that started with the original competency. This data will represent everything that is known about the extent, quality or quantity of the competency. Yet alone, it has no value and is uninterpretable; one must now look at this data and determine just how much of, or what degree of, the competency is or is not present. To do so requires a referent, a standard against which this data - an abstraction of performance - can be compared. Comparison of the data to such a standard enables one to base a decision on the data; to decide whether the student has achieved all, or part of, or none of, the desired competency. This standard will derive from judgment, and represents an a-priori decision of value. As such, it should be made public and known to all those on whom a decision about the competency bears.

The afore-mentioned and discussed steps do not, obviously, constitute a guide for establishment of a competency assessment program. To do so would require much more discussion about decision-making organization, learning management, and information systems. Neither do the steps discussed constitute a "how-to" guide. They are, in one respect, an elaborate apology for the lack of such a how-to guide - to my belief and knowledge, no such guide exists.

I have attempted first to discuss some of the things that competency assessment can do, and others that it cannot. I have placed it in a rather limited context of program evaluation, and defined it further as an evolving process. Some general criteria for competency assessment were recognized, and finally some basic steps for the construction of competency measures were defined and discussed. These steps are neither comprehensive nor theoretically founded; rather, they are meant to be practical and utilitarian. I believe that when they are followed, one can embark on a relatively sound and useful approach to assessing competencies in a teacher education program.

APPENDICES

Appendix A

NEEDS ASSESSMENT SURVEY
Plattsburgh Conference on
Assessment in Special Education CBTE Programs
August 11-13, 1976

Department of Professional Specialization
State University College
Plattsburgh, New York

CBTE Project
Division of Development
Support Services, Education
of Handicapped Children
State Education Department
Albany, New York

Please enter the name of your institution below:

_____

1. What benefits for your program would you like to obtain from this conference? Please indicate your interests by entering in the blanks:

   0 for "no interest" 1 for "slight interest"
   2 for "moderate interest" 3 for "great interest"

   Clearer identification of competency assessment needs in regard to:

   ____ a. what to assess

   ____ b. how extensively to assess

   ____ c. maximizing assessment validity

   ____ Knowledge of how other programs approach competency assessment

   ____ A pool of assessment items from which your program could draw

   ____ An improved system for collecting and managing competency assessment data

   ____ The development of criteria of competence

   ____ Other. Please specify.

   _____

   _____

   _____

2. Below are some general areas of competence. Please fill in the grid to indicate your program's current assessment status in regard to these areas.

AREA OF COMPETENCE	Degree of Need			Adequacy of Current Assessment			Check the Major Type of Current Assessment	
	None	Moderate	Great	Not Yet Attempted	Moderate	Very Good	Test	Observe
Managing problem behavior & affective environment								
Assessing learner characteristics, needs, progress								
Instructional planning skills								
Implementing instruction								
Interpersonal skills of program student								
Knowledge of exceptional children								
Other. Please specify.								

3. In regard to which setting do you wish to concentrate further efforts for developing competency assessment?

____ practicum ____ academics ____both

Appendix A

NEEDS ASSESSMENT SURVEY(continued)

4. Considering the interests expressed above, which of the following would be the most rewarding and productive use of your time at the conference?

0 for "none"
1 for "a little time"
2 for "a moderate amount of time"
3 for "a lot of time"

____ Large group presentations by experts

____ Small group presentations and discussions

____ Small group workshop activities with technical assistance

____ Other. Please Specify. _____

5. Please indicate the approximate amount of time you would like to spend with the following people by entering:

0 for "none
1 for "a little time"
2 for "a moderate amount of time"
3 for "a lot of time"

____ your own consortium

____ other consortia

____ teacher representatives

____school administrators

____ program faculty

____ consultants who can provide technical assistance

6. In which areas of exceptionality do you wish to concentrate your CBTE assessment efforts? Please check all applicable items in the grid below.

	Undergraduate Program	Graduate Program
Educable Mentally Retarded		
Trainable Mentally Retarded		
Emotionally Handicapped		
Physically Handicapped		
Multiply Handicapped		
Learning Disabled		
Speech & Hearing Impaired		
Blind & Vision Impaired		
Other. Please Specify. _____		

Appendix B

CONFERENCE PROGRAM

TUESDAY, AUGUST 10

3:00 p.m. - 6:00 p.m.	Registration and Orientation WHITEFACE HALL LOUNGE

WEDNESDAY, AUGUST 11

7:30 a.m. - 9:00 a.m.	Breakfast COLLEGE CENTER, BALLROOM A
8:30 a.m. - 10:00 a.m.	Final Registration COLLEGE CENTER, BALLROOM B
10:00 a.m. - 10:30 a.m.	Greetings from State University College at Plattsburgh, Bureau of Education for the Handicapped (USOE), Office for the Education of Children with Handicapping Conditions (SED) COLLEGE CENTER, BALLROOM B
10:30 a.m. - 12 NOON	Group interaction to clarify needs and goals of conference participants Dr. Theodore Andrews COLLEGE CENTER, BALLROOM B
12 NOON - 1:30 p.m.	Buffet lunch COLLEGE CENTER, BALLROOM A
1:30 p.m. - 4:30 p.m.	Workshop - Objective I: Developing or improving instruments for measuring student competencies other than interpersonal skills Dr. Gilbert Shearron Dr. James Strauch COLLEGE CENTER, BALLROOM B
4:30 p.m. - 5:30 p.m.	Staff Meeting
5:30 p.m. - 6:30 p.m.	Social Hour INFORMAL LOUNGE
6:30 p.m. - 7:30 p.m.	Dinner COLLEGE CENTER, BALLROOM A

Appendix B

CONFERENCE PROGRAM (continued)

THURSDAY, AUGUST 12

7:30 a.m. - 8:30 a.m.	Breakfast COLLEGE CENTER, BALLROOM A
9:00 a.m. - 12 NOON	Workshop - Objective II: Developing instruments to more adequately assess the interpersonal skills of students Dr. Paul Pottinger Dr. Joyce Spence COLLEGE CENTER, BALLROOM B
12 NOON - 1:30 p.m.	Lunch COLLEGE CENTER, BALLROOM A
1:30 p.m. - 4:30 p.m.	Workshop - Objective III: Specifying ways and means of insuring reliability and validity in competency assessment Dr. William Tikunoff Dr. Beatrice Ward Dr. Patricia Kay COLLEGE CENTER, BALLROOM B
4:30 p.m. - 5:30 p.m.	Staff Meeting
5:30 p.m. - 6:30 p.m.	Social Hour INFORMAL LOUNGE
6:30 p.m. - 7:30 p.m.	Dinner COLLEGE CENTER, BALLROOM A

FRIDAY, AUGUST 13

7:30 a.m. - 8:15 a.m.	Breakfast COLLEGE CENTER, BALLROOM A
8:30 a.m. - 11:30 a.m.	Workshop - Objective IV: Designing a system or components of a system for managing the assessment program Dr. Hugh Baird Dr. Robert Brinkerhoff COLLEGE CENTER, BALLROOM B
11:30 a.m. - 12 NOON	Conference Wrap-Up Mr. Joseph Gilmore COLLEGE CENTER, BALLROOM B
12 NOON	Lunch COLLEGE CENTER, BALLROOM A

Appendix C

CONFERENCE CONSULTANTS

Theodore E. Andrews
Andrews-Bryant Concern for
Educational Development
Washington, D.C.

J. Hugh Baird
Department of Secondary Education
Brigham Young University
Provo, Utah

Robert Brinkerhoff
Evaluation Research Center
University of Virginia
Charlottesville, Virginia

Patricia M. Kay
Competency Based Teacher
Education Project
Bernard M. Baruch College
City University of New York
New York, New York

Suzanne M. Kinzer
College of Education
University of Florida
Gainesville, Florida

Paul Pottinger
Institute for Competence
Assessment
McBer and Company
Boston, Massachusetts

Gilbert Shearron
Division of Elementary Education
University of Georgia
Athens, Georgia

Joyce Spence
The Child Center
Springfield, Oregon

James D. Strauch
Department of Educational
Psychology
University of Connecticut
Storrs, Connecticut

William J. Tikunoff
Far West Laboratory for Educational
Research and Development
San Francisco, California

Beatrice Ward
Far West Laboratory for Educational
Research and Development
San Francisco, California

Appendix D

CONFERENCE PARTICIPANTS

Robert Adrian
165 Cold Spray Road
Syosset, New York

Joan Benevento
400 East 58th Street
New York, New York

Cecile T. Benoit
215 West 83rd Street
New York, New York

Sister Anne Clark
College of Saint Rose
Albany, New York

Joseph Currier
56 Seneca Avenue
Dix Hills, New York

Rose Eells
20 Space Blvd. A
Albany, New York

John Finnerty
1488 East 18th Street
Brooklyn, New York

Kathy Foy
53 Main Street
Binghamton, New York

Robert Gaylord-Ross
Cory Gaylord-Ross
87 Mt. Bethel Road
Warren, New Jersey

Frank Gisonti
Torborg Gisonti
682 Tuckahoe Road
Yonkers, New York

Barry Glick
315 Cooper Lane
DeWitt, New York

Lyle Green
Colesville Road
Chenango Forks, New York

Ann Lee Halstead
36 Maple Street
Rouses Point, New York

Rick Kopp
215 Piedmont Avenue NE
Atlanta, Georgia

Mortimer Kreuter
Frances Kreuter
53-54 Bell Blvd.
Bayside, Long Island

Claire Lavin
Timothy O'Regan
95 North Little
New City, New York

Howard Margolis
Mtd. Rt. #8
Gravelly Point, Box 131
Plattsburgh, New York

Sister Maureen McCarthy
545 Broadway
Dobbs Ferry, New York

Sister Mary McCormick
Dominican College
Blauvelt, New York

Maryann McLaren
6 Woodbine Place
Delmar, New York

Agnes McNally
96 Prospect Street
Avon, New York

Dorothy S. Miller
Tiashoke, RD 1
Buskirk, New York

Appendix D

CONFERENCE PARTICIPANTS(continued)

Ann M. Mulholland
560 Riverside Drive
New York, New York

William Nadel
235 West 71st Street
New York, New York

Nedra Pawlewicz
80 Court Street
Plattsburgh, New York

Kathryn Pierce
543 South Mains
Canandaigua, New York

G. Mann Phillips
126 Andrew Avenue
Oakland, New Jersey

Ramon M. Rocha
5680 Main Street
Lakeville, New York

Paul Shaver
4 Hessian Drive
Schuylerville, New York

Sister Tina Marie Stocks
Dominican College
Blauvelt, New York

Maryann Sullivan
404 East 73rd Street
New York, New York

Sister Dorothy Wieting
Dominican College
Blauvelt, New York

Stan Wixson
94 Lincoln Avenue
Cortland, New York

Lois Wolf
255 West 88th Street
New York, New York

Appendix E

CONFERENCE PLANNING AND MANAGEMENT COMMITTEE

Theodore E. Andrews
Director
Andrews-Bryant Concern for Educational Development
702 Ninth Street S.E.
Washington, D.C. 20003

Joseph T. Gilmore
Director, CBTE Project in Special Education
The University of the State of New York
The State Education Department
Office for the Education of Children with Handicapping Conditions
55 Elk Street
Albany, New York 12224

Patricia M. Kay
Director, CBTE Project
Bernard M. Baruch College
School of Education
17 Lexington Avenue
New York, New York 10019

Suzanna M. Kinzer
Associate Professor
College of Education
University of Florida
Gainesville, Florida 32605

Raymond LaRocque
Coordinator of Educational Conferences
State University of New York
College of Arts and Sciences
Plattsburgh, New York 12901

Dale F. Nitzschke
Associate Dean, Faculty of Professional and General Studies
State University of New York
College of Arts and Sciences
Plattsburgh, New York 12901

William J. Penner
Director, Special Education CBTE Project
State University of New York
College of Arts and Sciences
Plattsburgh, New York 12901

Appendix F

CONFERENCE EVALUATION

Participants were asked to rate the conference on the following items. Mean ratings are given for each item.

Evaluation Item	Mean Rating
1. The organization of the conference was:	
poor 1 2 3 4 5 6 7 excellent	5.88
2. The objectives of the conference were:	
vague 1 2 3 4 5 6 7 clearly evident	5.89
3. The work of the consultants was:	
poor 1 2 3 4 5 6 7 excellent	6.17
4. The ideas and activities presented were:	
dull 1 2 3 4 5 6 7 very interesting	5.94
5. The scope was:	
inadequate 1 2 3 4 5 6 7 very adequate	5.94
6. My attendance at this conference should prove:	
of no benefit 1 2 3 4 5 6 7 very beneficial	6.39
7. Overall, I consider this conference to be:	
poor 1 2 3 4 5 6 7 excellent	6.00